Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Jakob Božič

Analiza katastrofalnega pozabljanja pri inkrementalnem učenju klasifikacijske konvolucijske nevronske mreže

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: izr. prof. dr. Danijel Skočaj

Ljubljana, 2019

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Ena izmed poznanih lastnosti umetnih nevronskih mrež je katastrofalno pozabljanje. Umetna nevronska mreža se v procesu učenja hitro prilagaja nazadnje podanim učnim primerom in dokaj hitro izgubi sposobnost modeliranja učnih podatkov, ki so v učnem procesu predstavljeni v zgodnji fazi, v kasnejših pa ne. To zelo otežuje implementacijo inkrementalnega učenja, kjer se učni primeri po uporabi zavržejo in se nadomestijo z novimi. Izkaže se, da nevronska mreža zelo uspešno modelira nove učne primere, starejših pa ne uspe več pravilno razpoznati. V diplomski nalogi podrobno analizirajte ta pojav na primeru problema klasifikacije slik s konvolucijsko nevronsko mrežo. Pokažite stopnjo katastrofalnega pozabljanja in raziščite vzroke, zaradi katerih se pojavi. Spremenite način osveževanja parametrov globoke nevronske mreže, s katerim naj bi upočasnili katastrofalno pozabljanje ter dosegli boljši klasifikacijski rezultat na testni množici, ki vsebuje učne primere tako iz predhodno, kot na novo naučenih razredov. Zasnujte in izvedite ustrezne eksperimente in rezultate primerno komentirajte.

Zahvaljujem se mentorju izr. prof. dr. Danijelu Skočaju za motivacijo in usmerjanje pri izdelavi diplomske naloge.

Laboratoriju za umetne vizualne spoznavne sisteme se zahvaljujem za dostop do ustrezne strojne opreme.

Zahvaljujem se tudi vsem ostalim, ki so pripomogli k nastanku tega dela in me podpirali v času študija.

Kazalo

Povzetek

Naslov: Analiza katastrofalnega pozabljanja pri inkrementalnem učenju klasifikacijske konvolucijske nevronske mreže

Avtor: Jakob Božič

Katastrofalno pozabljanje je pojav, ko umetna nevronska mreža ob inkrementalnem učenju novih nalog nemudoma in skoraj v celoti pozabi prejšnje. Problem je dobro znan in obstajajo različni pristopi k odpravljanju oz. omejevanju le-tega, vendar ga noben izmed pristopov ne reši v celoti.

V delu eksperimentalno preverimo, kateri so glavni dejavniki, ki privedejo do katastrofalnega pozabljanja. Analizo opravimo na globoki konvolucijski nevronski mreži, na problemu klasifikacije slik. Rezultate interpretiramo z matrikami zamenjav in grafi klasifikacijskih točnosti, spremembe uteži in odmikov tudi vizualiziramo. Dognanja iz analize uporabimo za zasnovo različnih pristopov k osveževanju parametrov, s katerimi želimo preprečiti oz. omiliti katastrofalno pozabljanje. Preverimo tudi scenarij, kjer imamo ob uporabi nevronske mreže na voljo oraklja, ki določi podmnožico razredov, v katere lahko klasificiramo primer. Implementiramo eno izmed obstoječih metod za odpravljanje katastrofalnega pozabljanja in jo prilagodimo, da deluje tudi brez oraklja.

Ugotovitve, predstavljene v delu, služijo kot izhodišče za zasnovo novih metod za odpravo katastrofalnega pozabljanja.

Ključne besede: katastrofalno pozabljanje, inkrementalno učenje, konvolucijske nevronske mreže, klasifikacija.

Abstract

Title: Analysis of catastrophic forgetting during incremental learning of classificational convolutional neural network

Author: Jakob Božič

Catastrophic forgetting is phenomenon when an artificial neural network immediately and almost completely forgets previously learned tasks when trained incrementally on new ones. It is a well-known problem and although there are many approaches to alleviating it, none of them solves it completely.

We experimentally check for main causes of catastrophic forgetting. Analysis is performed on a deep convolutional neural network for image classification. Results are interpreted by confusion matrices and classification accuracy graphs, we also visualize changes of weights and biases of network. Analytical findings serve as a basis for designing different approaches to updating network parameters, aiming to prevent or alleviate catastrophic forgetting. We also evaluate effects of availability of Oracle, capable of determining subset of all possible classes for classification, when using the network. We implement one of existing approaches to preventing catastrophic forgetting and adapt it to work without Oracle.

Findings, presented in thesis serve as a starting point for design of new approaches aimed at preventing catastrophic forgetting.

Keywords: catastrophic forgetting, incremental learning, convolutional neural networks, classification.

Poglavje 1

Uvod

1.1 Motivacija

Umetne nevronske mreže imajo v zadnjem času vse pomembnejšo vlogo pri razvoju umetne inteligence. Pojavljajo se na različnih področjih, od napovedovanja gibanja cen delnic na borzi do prepoznave govora in prevajanja, največji razcvet pa doživljajo na področju računalniškega vida. V zadnjem desetljetju se za probleme razpoznave objektov, segmentacije slik, sledenja itd. najbolj razvijajo globoke konvolucijske nevronske mreže, na določenih področjih pa že dosegajo in celo presegajo človeške zmožnosti.

Kljub hitremu razvoju se obstoječe nevronske mreže soočajo z raznimi težavami, od prevelikega prilagajanja podatkom iz učne množice (angl. overfitting) do potrebe po ogromnih količinah podatkov. Eden izmed odprtih problemov je tudi inkrementalno učenje globokih nevronskih mrež z izogibanjem katastrofalnega pozabljanja (angl. catastrophic forgetting).

Katastrofalno pozabljanje je pojav, ko nevronska mreža ob učenju na novih podatkih nemudoma in skoraj v celoti pozabi zakonitosti starih podatkov. K zasnovi in razvoju umetnih nevronskih mrež je deloma prispeval model bioloških nevronskih mrež, ki jih najdemo tudi v človeških možganih, vendar ljudje katastrofalnega pozabljanja ne izkušamo; npr. ko se naučimo vozit avtomobil, ne pozabimo, kako se vozi kolo. Rešitev problema katastrofalnega

pozabljanja bi pomembno prispevala k uporabnosti oz. zmožnosti uporabe istih mrež za različne probleme, ki se jih lahko naučimo inkrementalno.

1.2 Cilji diplomskega dela

V diplomskem delu želimo podrobno spoznati katastrofalno pozabljanje na domeni klasifikacije slik, njegove učinke in obseg ter kako hitro pride do le-tega. Predstavili bomo spreminjanje parametrov mreže, za katere domnevamo, da najbolj prispevajo k temu. Raziskali bomo, kateri so še ostali dejavniki, ki imajo vpliv na katastrofalno pozabljanje, in preverili, ali lahko z upoštevanjem le-teh omilimo katastrofalno pozabljanje. Ogledali si bomo matrike zamenjav in grafe klasifikacijskih točnosti, ki nazorno prikažejo obseg katastrofalnega pozabljanja.

Ko bomo imeli dober vpogled v dogajanje, ki privede do katastrofalnega pozabljanja, bomo poskušali to preprečiti oziroma vsaj zmanjšati ali upočasniti. Preverili bomo vpliv zamrznitve določenih parametrov oz. delov mreže, raziskali bomo, kaj se dogaja, če uporabimo različne optimizacijske metode, uporabili bomo različne stopnje učenja za različne dele mreže itd., implementirali bomo tudi enega izmed obstoječih pristopov k odpravaljanju katastrofalnega pozabljanja.

1.3 Sorodna dela

Problem katastrofalnega pozabljanja je bil podrobneje opisan že v poznih 80. letih prejšnjega stoletja [17, 18]. V [17] je predstavljeno katastrofalno pozabljanje na zelo preprostih polno povezanih nevronskih mrežah z le eno skrito plastjo. Avtorja predstavita, kaj se dogaja z nevronsko mmrežo, ki jo želita naučiti seštevanja dveh števil. Takoj je opazno, da mreža ob učenju drugega problema pozabi praktično vse znanje o prvem problemu. Katastrofalno pozabljanje poskusita preprečiti s spreminjanjem števila nevronov v skriti plasti, s spreminjanjem stopnje učenja in še nekaterimi pristopi, ven-

dar neuspešno.

Avtor v [18] ravno tako raziskuje problem na polno povezanih nevronskih mrežah z le eno skrito plastjo. Opiše, kaj se dogaja, če ob učenju na novem problemu spreminjamo vse uteži mreže, kaj se dogaja, če spreminjamo le tiste, ki so se na prvotnem problemu najmanj spreminjale, kakšen vpliv ima dodajanje novih nevronov v skriti plasti in kaj se dogaja, če spreminjamo vse uteži ali le tiste, ki pripadajo novim nevronom. Avtor sklene, da noben izmed pristopov ne rešuje problema katastrofalnega pozabljanja.

Prvi pristopi, ki so zmanjševali katastrofalno pozabljanje v omejenih okoliščinah, so se pojavili kmalu zatem. V [6] avtor trdi, da pride do katastrofalnega pozabljanja zaradi prekrivajočih se aktivacij v skriti plasti, tj. zaradi deljenih predstavitev. Predlaga algoritem „ostrenja vozlišč“ (angl. node sharpening), ki zmanjšuje prekrivanje s povečevanjem aktivacije najbolj aktivnih nevronov v skriti plasti in zmanjševanjem aktivacije najmanj aktivnih. Tako nevronsko mrežo prisili, da se aktivacije v skriti plasti med različnimi primeri čim manj prekrivajo med sabo oz. so čim bolj pravokotne ena na drugo.

Eden izmed pristopov, ki se ravno tako kot „ostrenje vozlišč“ zanaša na ortogonalizacijo, je predstavljen v [5]. Pristop se ravno tako zanaša na redke (angl. sparse) predstavitve vhodnih podatkov, kar seveda ne pride v poštev pri delu s slikami.

Pojav grafičnih kartic, ki so z mnogo računskimi enotami zelo pohitrile učenje [4], razvoj boljših algoritmov in metod za preprečevanje prenasičenja (angl. overfitting) [11, 20] in obsežne količine podatkov so privedli do popularizacije globokih (konvolucijskih) nevronskih mrež. Avtorji v [7] preverijo, kako na katastrofalno pozabljanje vpliva izbira aktivacijske funkcije, kakšen vpliv ima uporaba izpadne plasti (angl. Dropout) in kako na katastrofalno pozabljanje vpliva velikost nevronske mreže oz. število parametrov le-te.

V zadnjih letih so se pojavile nekatere metode, ki do določene mere rešujejo problem katastrofalnega pozabljanja, vendar se za razliko od prvotnih ne zanašajo na prilagajanje predstavitev v skritih plasteh.

Avtorji v [13] spremenijo kriterijsko funkcijo tako, da upočasnuje spreminjanje parametrov, ki so bolj pomembni za prej naučene probleme. Ob učenju na novem problemu se aproksimira pomembnost parametrov na prejšnjih in kaznuje spremembe pomembnejših.

Eden izmed pristopov je tudi prilaganje kriterijske funkcije, da kaznuje take spremembe parametrov, ki povzročajo večje padce v natančnostih na prej naučenih podatkih. Avtorji v [3] predlagajo uporabo kriterijske funkcije, ki združuje križno entropijo (angl. cross entropy) in distilacijsko funkcijo, predstavljeno v [9]. Za izračun distilacijske funkcije moramo sicer ohraniti del prejšnje učne množice. Rezultati tega pristopa so trenutno med najboljšimi.

Nekateri izmed pristopov zahtevajo, da ob uporabi nevronske mreže za posamezni primer vemo, v katere razrede lahko le-tega razvrstimo. Torej če smo učili mrežo v npr. treh fazah, moramo vedeti, ali ta primer pripada razredom, ki smo se jih naučili v prvi, v drugi ali v tretji fazi. V praksi to sicer ni vedno mogoče, poleg tega je pa ta problem že v osnovi lažji.

Maskiranje uteži [16] je eden izmed takih pristopov. Za učenje prve naloge se nauči nevronsko mrežo oz. se douči obstoječo, nato pa za vsako novo nalogo pridobimo binarno masko, ki za vsak parameter določa, ali ostane nespremenjen ali se nastavi na 0. Mreža se tako s katastrofalnim pozabljanjem sploh ne sooča, saj so vrednosti parametrov neodvisne od števila naučenih nalog. Slabost te metode je, poleg tega da moramo vedeti, katero masko uporabiti za kateri problem, tudi ta, da moramo za vsako naučeno nalogo shraniti binarno masko celotne mreže, kar lahko hitro prevede do ogromnih količin podatkov za shranjevanje.

Avtorji metode „spominsko občutljive povezave“ (angl. Memory Aware Synapses) [1] ravno tako predlagajo modificiranje kriterijske funkcije, tako da se v poznejših fazah učenja bolj pomembni parametri manj spreminjajo. Prednost te metode je, da lahko občutljivost parametrov izračuna na katerikolih primerih, tudi neoznačenih. Ob koncu učenja v določeni fazi se izračuna pomembnost vseh parametrov, ob učenju pa se te pomembnosti uporabi kot regularizacijski del v kriterijski funkciji. Čeprav avtorji v eksperimentalnem

delu privzamejo, da vemo za vsak testni primer, kateri podmnožici možnih razredov pripada, se metodo lahko prilagodi, da ta predpostavka ni več potrebna.

1.4 Prispevki

Glavni prispevek diplomskega dela je podrobna analiza katastrofalnega pozabljanja v glokokih konvolucijskih nevronskih mrežah za klasifikacijo slik. Kvantitativno bomo predstavili obseg in hitrost katastrofalnega pozabljanja. Vizualizirali bomo spremembe parametrov nevronskih mrež, za katere lahko sklepamo, da imajo največji vpliv na pojav katastrofalnega pozabljanja. Izvedeni eksperimenti bodo z različnimi pristopi k osveževanju parametrov mreže pripomogli k boljšemu razumevanju katastrofalnega pozabljanja.

Drugi prispevek dela je primerjava različnih načinov osveževanja parametrov nevronske mreže, npr. zamrznitev ali spremenjena stopnja učenja, oz. njihovega vpliva na pojav katastrofalnega pozabljanja. Implementirali bomo tudi enega izmed obstoječih pristopov za zmanjševanje katastrofalnega pozabljanja, ki se zanaša na obstoj oraklja ob uporabi nevronske mreže. Isti pristop bomo prilagodili, da bo deloval v standardnem okolju, kjer orakelj ni na voljo.

1.5 Struktura dela

V Poglavju 2 bomo predstavili delovanje umetnih nevronskih mrež. Spoznali bomo, kako le-te delujejo, kateri so osnovni gradniki, kakšne parametre imajo le-te, kaj so aktivacijske in kaj kriterijske funkcije. Nato bomo opisali, kako se nevronske mreže uči z metodo vzvratnega razširjanja. Poleg preprostih, polno-povezanih nevronskih mrež, bomo spoznali tudi nekaj lastnosti globokih konvolucijskih nevronskih mrež, ki so trenutno glavni akterji na področju klasifikacije slik.

V Poglavju 3 bomo najprej podrobno opisali, kaj sploh je inkremen-

talno učenje in kaj je katastrofalno pozabljanje. Nato bomo predstavili preizkušene načine zmanjševanja katastrofalnega pozabljanja, večina jih temelji na različnih shemah osveževanja parametrov, ki jih bomo izvedli v Poglavju 4. Predstavili bomo tudi pristop k učenju z orakljem, ki predpostavlja, da lahko ob testiranju mreže omejimo nabor možnih razredov za klasifikacijo. Spoznali bomo enega izmed pristopov za dušenje katastrofalnega pozabljanja in ga tudi implementirali, tako z uporabo oraklja kot tudi brez. Na koncu bomo še predstavili arhitekturo konvolucijske nevronske mreže, ki se uporablja v vseh eksperimentih.

Poglavje 4 vsebuje eksperimente, ki nam v praksi pokažejo katastrofalno pozabljanje in poskuse k odpravljanju le-tega. Vizualizirali bomo spremembe parametrov nevronske mreže, ki pripeljejo do katastrofalnega pozabljanja. Preverili bomo, kako različni načini osveževanja parametrov vplivajo na obseg in hitrost katastrofalnega pozabljanja. Predstavili bomo rezultate vseh izvedenih eksperimentov in jih ovrednotili.

V Poglavju 5 bomo predstavili sklepne ugotovitve diplomskega dela in predstavili priložnosti na nadaljnje raziskovanje.

Poglavje 2

Umetne nevronske mreže

Umetne nevronske mreže so računski modeli, zgrajeni iz umetnih nevronov, ki temeljijo na delovanju bioloških nevronov. Poznamo več vrst umetnih nevronskih mrež, najpreprostejše so polno povezane mreže (angl. fully connected neural network), na področju računalniškega vida imajo najpomembnejšo vlogo konvolucijske nevronske mreže (angl. convolutonal neural networks), za prevajanje in obdelavo jezika (angl. natural language processing) se največ uporabljajo rekurenčne nevronske mreže (angl. recurrent neural networks), za generiranje umetnih podatkov pa se uporabljajo generativne nevronske mreže (angl. generative adversarial networks), obstajajo pa tudi druge vrste umetnih nevronskih mrež. Vse umetne nevronske mreže si delijo osnovne gradnike, ki jih bomo podrobneje predstavili, nato pa bomo pobliže pogledali še arhitekturo in posebnosti konvolucijskih nevronskih mrež.

2.1 Struktura

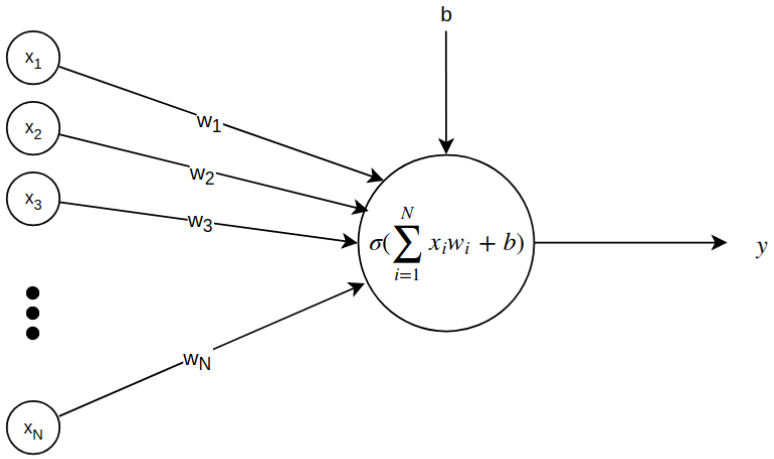
Umetne nevronske mreže so sestavljene iz umetnih nevronov, združenih v plasti. Umetni nevron je preprost gradnik, ki temelji na poznavanju bioloških nevronov, ki ima le to funkcijo, da vhode preslika v izhod. Vhod predstavimo z množico izhodov iz nevronov prejšnje plasti $X = \{x_i\}$, kjer je i indeks vhoda oz. indeks nevrona iz prejšnje plasti kateremu pripada. Vsak nevron

ima svojo množico uteži $W = w_i$, ki določa, kako se utežijo posamezni vhodi, i predstavlja indeks vhoda. Izhod nevrona določa tudi odmik (angl. bias) b, ki se prišteje uteženi vsoti pred transformacijo z aktivacijsko funkcijo σ. Izhod nevrona lahko torej zapišemo kot

$$a = \sigma\Big(\sum_{i=1}^{N} x_i * w_i + b\Big), \tag{2.1}$$

kjer N predstavlja število nevronov v prejšnji plasti, oz. v vektorski obliki

$$a = \sigma\Big(XW^T + b\Big). \tag{2.2}$$



Slika 2.1: Delovanje umetnega nevrona

Umetne nevronske mreže so sestavljene iz plasti, te pa iz več nevronov, zato moramo v zgornjih enačbah dodati še nekaj indeksov. Plasti številčimo od vhoda v mrežo proti izhodu, vhodna plast ima torej indeks 1, izhodna pa L, kjer je L število vseh plasti. X^{l-1} predstavlja vektor izhodov iz plasti $l-1$, tj. vhodov v plast l. Matrika $W^l = \{w^l_{j,k}\}$ predstavlja uteži vseh nevronov plasti l, posamezen element je utež za vhod k-tega nevrona iz plasti $l-1$ v j-ti nevron v plasti l. Vektor $b^l = \{b^l_j\}$ predstavlja odmike (angl. biases)

nevronov v plasti l. Izhode iz plasti l a^l lahko torej preprosto in učinkovito izračunamo kot

$$a^l = \sigma\Big(W^l a^{l-1} + b^l\Big). \tag{2.3}$$

2.1.1 Aktivacijske funkcije

Aktivacijska funkcija σ preslika vsoto odmikov in utežene vsote vhodov v aktivacijo, tj. izhod nevrona. Aktivacijska funkcija vnaša v nevronsko mrežo nelinearno preslikavo, kar mreži omogoča, da se nauči aproksimacije katerekoli kompleksne funkcije. Označimo vmesno količino kot

$$z = \sum_{i=1}^{N} x_i w_i + b. \tag{2.4}$$

Izhod (aktivacija) nevrona je torej definiran kot

$$a = \sigma(z). \tag{2.5}$$

Vrednost z lahko zavzame zelo velike ali zelo majhne vrednosti, kar pa v nevronskih mrežah ni zaželjeno. Aktivacijske funkcije te vrednosti omejijo, nekatere na določen interval, nekatere odrežejo negativne vrednosti, spet druge lahko izhod binarizirajo. Predstavili bomo najbolj uporabljene funkcije.

Sigmoidna funkcija

Sigmoidna funkcija preslika vhod na interval $[0, 1]$. Določena je kot

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{2.6}$$

Dolgo časa je bila sigmoidna funkcija najbolj pogosto uporabljena aktivacijska funkcija, vendar jo v zadnjem času pogosteje nadomeščajo druge, tako zaradi boljših rezultatov kot zaradi hitrejšega izračuna.

ReLU

Funkcija ReLU (angl. rectified linear unit) odreže negativne vrednosti vhoda, definirana je kot

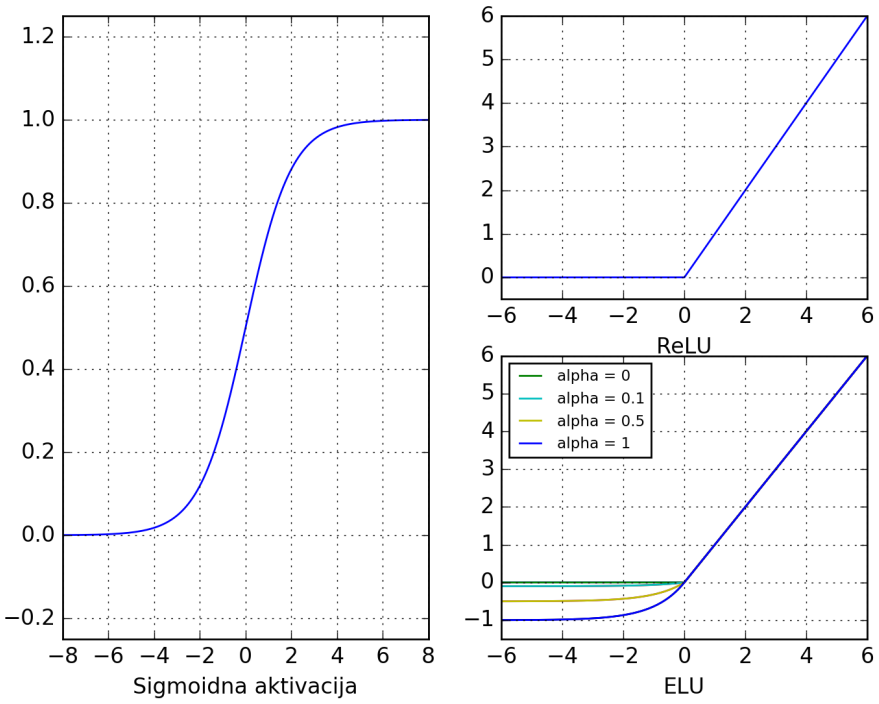
$$\sigma(x) = max(0, x). \tag{2.7}$$

Zaradi svoje preprostosti in dobrih rezultatov se uporablja zelo pogosto, predvsem v globokih nevronskih mrežah.

ELU

Funkcija ELU je definirana s parametrom α, izhod omeji na interval $[-\alpha, \infty]$. Definirana je kot

$$\sigma(\alpha, x) = max(0, x) + min(0, \alpha * (e^x - 1)). \tag{2.8}$$



Slika 2.2: Aktivacijske funkcije

Softmax

Softmax je funkcija, ki se uporablja za pretvorbo izhodov iz zadnje plasti nevronske mreže v verjetnostno porazdelitev. Od ostalih aktivacijskih funkcij

se razlikuje po tem, da potrebuje za izračun vse izhode iz plasti. Posamezna komponenta je definirana kot

$$\sigma(x_j) = \frac{e^{x_j}}{\sum_{i=1}^{N} e^{x_i}}, \tag{2.9}$$

kjer je N število izhodnih nevronov, x_i pa predstavlja aktivacijo i-tega nevrona v zadnji plasti pred transformacijo.

2.1.2 Kriterijske funkcije

Kriterijska funkcija (angl. cost function, tudi loss function) nam pove, kako daleč od pravilne rešitve je izhod iz mreže. Izpolnjevati mora dva pogoja: da jo lahko izračunamo kot povprečje vrednosti za posamezne učne primere ter da jo lahko zapišemo kot funkcijo izhodov nevronske mreže. Prvi pogoj je potreben ker se gradient funkcije izračuna za vsak učni primer posebej, moramo pa ga posplošiti na celotno funkcijo. Drugi pogoj zagotavlja, da lahko napako v zadnji plasti mreže razširjamo nazaj po mreži.

Srednja kvadratna napaka

Srednja kvadratna napaka (angl. mean squared error) meri povprečno odstopanje izhoda od resnične vrednosti. Definirana je kot

$$MSE(Y, \hat{Y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i), \tag{2.10}$$

kjer N predstavlja število izhodnih nevronov, y_i predstavlja izhod i-tega nevrona, $\hat{y}_i$ pa resnično vrednost istega nevrona.

Križna entropija

Križno entropijo (angl. cross entropy) lahko uporabljamo, ko je naš izhod iz mreže verjetnostna porazdelitev; izračuna razliko med napovedano in resnično verjetnostjo razreda. V kolikor izhod iz naše mreže ni verjetnosta porazdelitev, ga lahko pretvorimo v le-to z uporabo aktivacijske funkcije

Softmax. Definirana je kot

$$CC(Y, \hat{Y}) = -\sum_{i=1}^{N} y_i \log \hat{y}_i. \tag{2.11}$$

2.2 Učenje

Na umetne nevronske mreže lahko gledamo kot na matematične funkcije, ki nam vhod, parametre in resnični izhod (angl. ground truth) preslikajo v skalarno vrednost: izgubo, vrednost kriterijske funkcije. Nad vhodom in resničnim izhodom nimamo nadzora, zato ju ne obravnavamo kot parametra funkcije. Učenje nevronskih mrež torej predstavlja minimizacijo vrednosti kriterijske funkcije s prilagajanjem vrednosti parametrov.

Vrednosti parametrov spreminjamo glede na gradient funkcije, predpogoj za izračun gradienta pa so vrednosti parcialnih odvodov funkcije po posameznih parametrih. Za izračun parcialnih odvodov obstaja več pristopov, analitičen izračun ali numerična aproksimacija sta v praksi neizvedljiva, obstoječ način učenja nevronskih mrež pa je metoda vzvratnega razširjanja (angl. backpropagation).

Ko poznamo gradient oz. vrednosti parcialnih odvodov funkcije po posameznih parametrih, uporabimo eno izmed optimizacijskih metod, ki temeljijo na gradientnem spustu.

2.2.1 Pripava podatkov

Za uspešno učenje nevronskih mrež je potrebna velika količina podatkov. Pred učenjem je potrebno množico podatkov disjunktno ločiti na vsaj učno in testno množico, včasih pa želimo imeti tudi tretjo, validacijsko množico. Zelo pomembno je, da se mreže ne testira na podatkih, na katerih je bila naučena, saj so rezultati v tem primeru pogosto zavajajoči.

2.2.2 Optimizacijske metode

Optimizacijske metode prilagajajo vrednosti parametrov Θ nevronske mreže na podlagi gradienta. Velika večina metod sprejme kot parameter stopnjo učenja (angl. learning rate) η, nekatere pa še dodatne parametre.

SGD

Stohastični gradienti spust (angl. stochastic gradient descent) je najpreprostejša izmed optimizacijskih metod, ki jih uporabljamo za učenje nevronskih mrež. Edini parameter, ki ga sprejme, je stopnja učenja η. Parametre Θ nevronske mreže prilagodi po formuli

$$\theta_i = \theta_i - \eta \frac{\partial C}{\partial \theta_i}, \tag{2.12}$$

kjer $\frac{\delta C}{\delta \theta_i}$ predstavlja parcialni odvod kriterijske funkcije C po parametru θ_i.

Adam

Adaptive moment estimation [12] je optimizacijska metoda, ki upošteva drseče povprečje m_t (prvi moment) in varianco v_t (drugi moment) gradienta. Drseče povprečje in varianco posameznega parametra izračuna kot

$$\begin{aligned} m_i^t &= \beta_1 m_i^{t-1} + (1-\beta_1)\frac{\partial C}{\partial \theta_i} \\ v_i^t &= \beta_2 v_i^{t-1} + (1-\beta_2)\Big(\frac{\partial C}{\partial \theta_i}\Big)^2, \end{aligned} \tag{2.13}$$

β_1 in β_2 sta hiperparametra, njuni privzeti vrednosti sta 0,9 in 0,999 in ju zelo redko spreminjamo. Ker sta zgornji formuli pristranski oceni za prvi in drugi moment, moramo pristranskost odpraviti (angl. bias correction)

$$\begin{aligned} \hat{m}_i^t &= \frac{m_i^t}{1-\beta_1^t} \\ \hat{v}_i^t &= \frac{v_i^t}{1-\beta_2^t} \end{aligned} \tag{2.14}$$

Ostane samo še posodobitev parametrov Θ

$$\theta_i^t = \theta_i^{t-1} - \eta \frac{\hat{m}_i^t}{\sqrt{\hat{v}_i^t} + \epsilon} \tag{2.15}$$

kjer je η stopnja učenja, ϵ pa poljubno majhno število, da se izognemo deljenju z 0. Adam se zadnje čase vse pogosteje uporablja, saj konvergira hitreje kot SGD.

2.2.3 Metoda vzvratnega razširjanja

Metoda vzvratnega razširjanja (angl. backpropagation) predstavlja učinkovit način za izračun gradienta kriterijske funkcije nevronske mreže. Metoda se zanaša na verižno pravilo (angl. chain rule) za iteativni izračun gradientov za vsako plast nevronske mreže.

Verižno pravilo

Verižno pravilo nam omogoča izračun odvoda kompozituma dveh ali več funkcij. V kolikor je z odvisen od y, le-ta pa od x, potem velja

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}. \tag{2.16}$$

Gradient kriterijske funkcije

Za izračun gradienta kriterijske funkcije moramo izračunati parcialne odvode kriterijske funkcije C po vseh njenih parametrih, tj. utežeh in odmikih $\frac{\partial C}{\partial \theta_i}$.

Parcialni odvod kriterijske funkcije C glede na izhod (aktivacijo) i-tega nevrona v l-ti plasti označimo označimo z

$$\delta_i^l = \frac{\partial C}{\partial a_i^l}. \tag{2.17}$$

Vemo, da velja $a_i^l = \sigma(z_i^l)$, torej po verižnem pravilu sledi

$$\begin{aligned} \frac{\partial C}{\partial z_i^l} &= \frac{\partial C}{\partial a_i^l} \cdot \frac{\partial a_i^l}{\partial z_i^l} \\ &= \delta_i^l \cdot \sigma'(z_i^l). \end{aligned} \tag{2.18}$$

Parcialne odvode kriterijske funkcije glede na odmike torej izračunamo kot

$$\begin{aligned} \frac{\partial C}{\partial b_i^l} &= \frac{\partial C}{\partial a_i^l} \cdot \frac{\partial a_i^l}{\partial z_i^l} \cdot \frac{\partial z_i^l}{\partial b_i^l} \\ &= \delta_i^l \cdot \sigma'(z_i^l) \cdot 1, \end{aligned} \tag{2.19}$$

glede na uteži pa kot

$$\begin{aligned}\frac{\partial C}{\partial w_{i,k}^l} &= \frac{\partial C}{\partial a_i^l} \cdot \frac{\partial a_i^l}{\partial z_i^l} \cdot \frac{\partial z_i^l}{\partial w_{i,k}^l} \\ &= \delta_i^l \cdot \sigma'(z_i^l) \cdot a_k^{l-1}.\end{aligned} \tag{2.20}$$

Ostane nam torej le še izračun δ_i^l, ki se razlikuje glede na to, ali gre za zadnjo plast v nevronski mreži ($l = L$) ali pa za vmesno. V kolikor gre za zadnjo plast, je izračun odvisen od tega, katero kriterijsko funkcijo smo izbrali. Za preostale plasti velja

$$\begin{aligned}\frac{\partial C}{\partial a_i^{l-1}} &= \sum_{\forall k} \Big(\frac{\partial C}{\partial a_k^l} \cdot \frac{\partial a_k^l}{\partial z_k^l} \cdot \frac{\partial z_i^l}{\partial a_k^{l-1}} \Big) \\ &= \sum_{\forall k} \Big(\delta_i^l \cdot \sigma'(z_i^l) \cdot q_{i,k}^l \Big),\end{aligned} \tag{2.21}$$

kjer k predstavlja indeks nevronov v plasti l. Sedaj imamo vse, kar potrebujemo za izračun gradienta kriterijske funkcije. Poleg učne množice U, kriterijske funkcije C in optimizacijske metode O potrebujemo še število epoh učenja nE in velikost paketov, na katere razdelimo učno množico vP. Celoten postopek učenja nevronske mreže torej izgleda:

- za število epoh nE ponavljaj:
  - premešaj in loči učno množico U na pakete p velikosti vP
  - za vsak paket p izvedi:
    - inicializiraj parcialne odvode uteži dW in odmikov dB na 0
    - za vsak učni primer u iz paketa p izvedi:
      - izračunaj izhod mreže y, shrani z_i^l za vsak nevron
      - izračunaj parcialne odvode zadnje plasti $\frac{\partial C}{\partial a_i^L}$ glede na resnični izhod y in kriterijsko funkcijo C
      - od predzadnje proti prvi plasti izračunaj parcialne odvode (2.21)
      - izračunaj parcialne odvode uteži (2.20) in odmikov (2.19) ter jih prištej dW in dB

* dW in dB deli z vP
* z optimizacijsko metodo O prilagodi vrednosti uteži in odmikov glede na dW in dB

2.2.4 Globoke konvolucijske nevronske mreže

Na področju umetnega zaznavanja se v zadnjem času največ uporablja globoke konvolucijske nevronske mreže. Dosegajo boljše rezultate, saj upoštevajo lokalnost vizualnih podatkov, poleg tega imajo zaradi deljenja parametrov manj le-teh kot polno povezane mreže. Diskretna konvolucija dvodimenzionalnega vhoda I s filtrom K je definirana kot

$$S(i,j) = (K * I)(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i-m, j-n)K(m,n), \tag{2.22}$$

kjer sta M in N širina oz. višina konvolucijskega filtra. Kljub temu večina ogrodij za delo z nevronskimi mrežami uporablja križno korelacijo (angl. Cross Corelation), ki je definirana zelo podobno

$$S(i,j) = (I * K)(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i+m, j+n)K(m,n). \tag{2.23}$$

Opazimo, da je konvolucija dejansko križna korelacija s filtrom, zrcaljenim preko obeh dimenzij. Večina ogrodij za delo z nevronskimi mrežami dejansko uporablja križno korelacijo, saj se mreža filtrov nauči in se tako ob križni korelaciji mreža nauči zrcaljenih filtrov, ki dajo enake rezultate kot konvolucija.

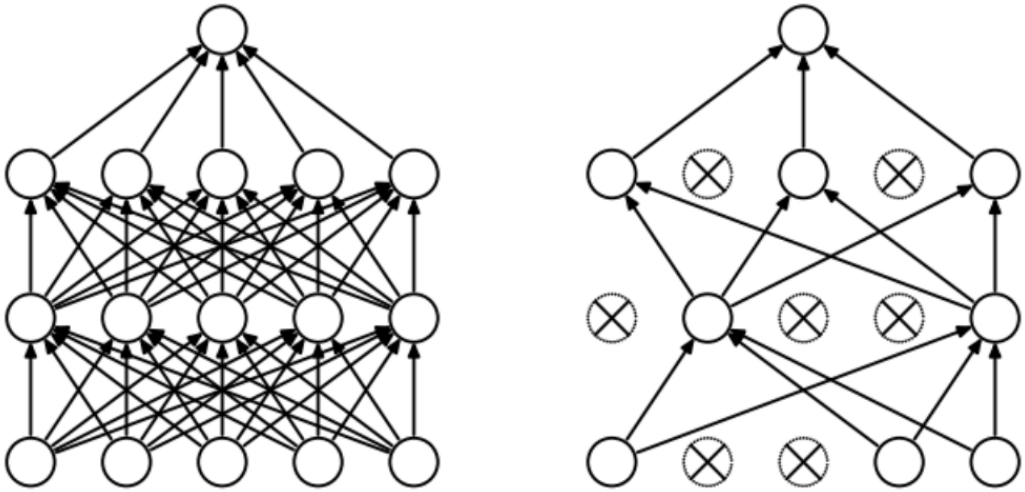
Konvolucijske plasti

Konvolucijske plasti izvajajo operacije konvolucije. Vhod v konvolucijsko plast je dimenzije $W \times H \times D$. Izhod je odvisen od dimenzij filtra $M \times N \times D$, števila naučenih filtrov N, koraka (angl. stride) p, r in od tega, ali uporabljamo dodajanje obrobe (angl. padding) ali ne. Za vsak naučen filter izvedemo konvolucijo po celotnenm vhodu. Če uporabljamo dodajanje obrobe, se prvi dimenziji vhoda zmanjšata za faktor koraka (oz. ohranita, če je

korak 1), drugače moramo korak in dimenzije filtra ustrezno upoštevati pri zasnovi plasti. V kolikor uporabljamo dodajanje obrobe, so izhodne dimenzije $\frac{W}{p} \times \frac{H}{r} \times N$. Konvolucijska plast se nauči filtrov (uteži) in odmikov za vsak filter.

Izpadna plast

Izpadna (angl. Dropout) plast [20] preprečuje oz. omejuje preveliko prilagajanje učnim podatkom. Med učenjem mreže nevron izpade z verjetnostjo p,



Slika 2.3: Izpadna plast, levo nevronska mreža pred izpadom posameznih nevronov, desno po izpadu. Povzeto po [20].

prav tako izpadejo vse njegove vhodne in izhodne povezave. Izpad naključnih nevronov prinaša podobne rezultate, kot če bi imeli opravka z 2^n (n predstavlja število nevronov, ki lahko izpadejo) mrežami, kar prinaša bolj robustne značilke. Ob uporabi mreže se upoštevajo vsi nevroni, njihove aktivacije pa se množijo s faktorjem p, tako da se upošteva izpad v fazi učenja.

Paketna normalizacija

Paketna normalizacija [11] (angl. Batch Normalization) omogoča boljše rezultate zaradi odpornosti na premik kovariance (angl. covariance shift), po-

sledično dovoljuje večje stopnje učenja, prav tako pa sčiti pred prevelikim prilaganjem učnim primerom. Ko učimo nevronske mreže, pogosto normaliziramo učno množico, tj. odštejemo povprečno vrednost μ in delimo z odklonom σ

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.24}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2.25}$$

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2}}. \tag{2.26}$$

Paketna normalizacija opravlja podobno funkcijo med plastmi mreže, saj se lahko vhodi v posamezne plasti med različnimi učnimi primeri zelo razlikujejo. Plast paketne normalizacije se nauči le dva parametra, γ in β, ki vplivata na izhod iz plasti po enačbi

$$y_i = \gamma\hat{x}_i + \beta \tag{2.27}$$

Med učenjem mreže se povprečno vrednost μ in odklon σ izračuna na paketu učnih podatkov (angl. batch), med uporabo pa se upošteva povprečje vseh vrednosti iz faze učenja.

Združevalne plasti

Združevalne (angl. pooling) plasti se uporabljajo za zmanjševanje dimenzij podatkov med plastmi nevronske mreže. Delujejo tako, da vrednosti izhodov iz enega območja združijo v eno vrednost, velikost območja je parameter plasti. Med najbolj uporabljenimi sta združevanje z maksimizacijo (angl. max pooling) in združevanje s povprečenjem (angl. average pooling). Združevanje z maksimizacijo vse vrednosti iz enega območja združi v največjo izmed teh vrednosti, združevanje s povprečenjem pa v povprečno vrednost le-teh. V praksi daje združevanje z maksimizacijo boljše rezultate, alternativa obema pa je lahko kovolucijska plast z večjim korakom.

Poglavje 3

Katastrofalno pozabljanje

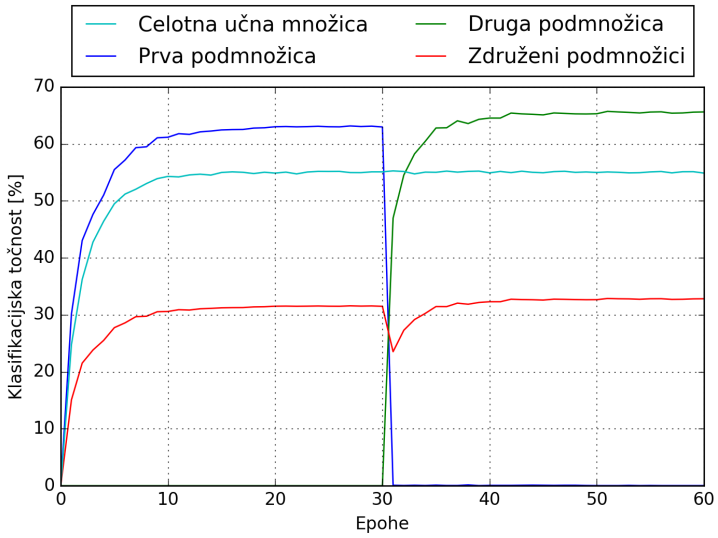
Sodobne globoke umetne nevronske mreže imajo lahko tudi po več deset milijonov parametrov in lahko tako zavzamejo veliko prostora, kar včasih predstavlja problem, predvsem na mobilnih in ostalih napravah z omejenim prostorom. Nič neobičajnega ni, da so podatkovne zbirke, na katerih se nevronske mreže učijo, velikosti več deset gigabajtov, učenje mreže od začetka pa lahko kljub večjemu številu grafičnih kartic traja več ur ali celo dni. V zadnjem času se vse več poudarka daje tudi varstvu osebnih podatkov, prav tako se ljudje in podjetja zavedajo vrednosti podatkov samih, zato obstaja možnost, da nam podatki po učenju niso več na voljo.

Eden izmed pristopov k obema problemoma je inkrementalno učenje nevronskih mrež. Ob prejetju novih učnih podatkov in morebitni predpostavki, da nam stari niso več na voljo, želimo obstoječo nevronsko mrežo doučiti na novih podatkih v čim krajšem možnem času; učenja od začetka ne želimo, saj vzame preveč časa, če nam stari podatki niso na voljo, pa niti ni možno.

V idealnem scenariju bi nam inkrementalno učenje na novih učnih podatkih omogočalo, da dosežemo enako dobre rezultate, kot če bi imeli že na začetku na voljo vse podatke. V teoriji ne bi bilo razlike med tem, ali mrežo naučimo na npr. 100 tisoč učnih primerih razdeljenih v 100 razredov ali mrežo najprej naučimo na 50 tisoč učnih primerov iz 50 razredov, nato te podatke zavržemo in mrežo doučimo na drugih 50 tisoč primerih razdelje-

nih v drugih 50 razredov. Realnost je nažalost daleč od tega, saj pride do problema katastrofalnega pozabljanja.

Katastrofalno pozabljanje [17] je fenomen, ko mreža ob dodajanju novih razredov in učenju na novih podatkih hitro in skoraj v celoti pozabi zakonitosti prejšnjih.



Slika 3.1: Klasifikacijske točnosti pri inkrementalnem učenju.

Graf na Sliki 3.1 prikazuje, kaj se dejansko dogaja v zgoraj opisanem scenariju. Svetlo modra črta prikazuje klasfikacijsko točnost v primeru, da imamo že na začetku na voljo vse učne podatke. Temno modra in zelena črta prikazujeta klasifikacijske točnosti ob učenju na dveh podmnožicah, vsaka vsebuje polovico učnih podatkov in različne razrede. Rdeča črta prikazuje klasifikacijsko točnost na vseh testnih podatkih ob učenju na ločenih podmnožicah. Takoj lahko vidimo, da klasifikacijska točnost na prvi podmnožici nemudoma pade na 0% ob začetku učenja na drugi podmnožici, posledično se rdeča črta nikoli niti najmanj ne približa svetlo modri črti, želeli bi, da se združita.

Eksperimentalno bomo raziskali, zakaj pride do tako izrazitega pozabljanja in kako, če sploh, se mu lahko izognemo. Zanima nas, kako se spreminjajo uteži in odmiki v zadnji plasti in ali lahko te spremembe vsaj deloma pojasnijo, zakaj pride do pozabljanja. Matrike zamenjav nam bodo pomagale razumeti, kako mreža uvrsti testne primere, ali jih večinoma narobe razvrsti v razrede iz iste faze učenja ali iz druge.

3.1 Zasnova eksperimentov

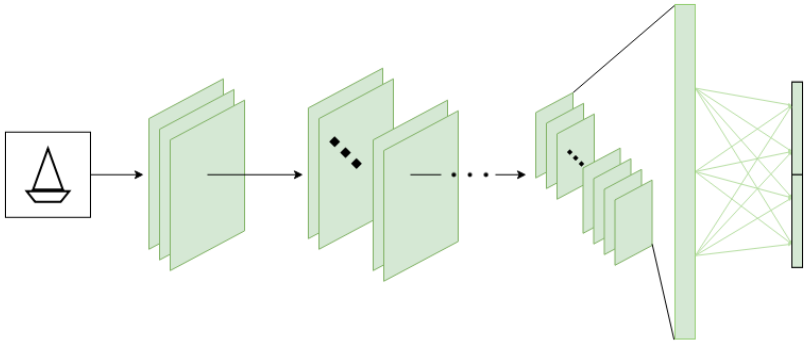
3.1.1 Klasični scenarij

Najprej bomo predstavili eksperimente, v katerih ob uporabi mreže nimamo na voljo oraklja in ne moremo omejiti množice razredov, v katere lahko klasificiramo primer.

3.1.1.1 Naivni pristop

Najprej si bomo pogledali, kako sploh pride do katastrofalnega pozabljanja. Podrobno bomo raziskali, kaj se dogaja s parametri mreže v prvi epohi druge faze učenja, v kateri je katastrofalno pozabljanje načeloma že zelo izrazito. Preverili bomo, kaj se dogaja, če dovolimo spremembe vseh parametrov nevronov v zadnji plasti, ne glede na to ali pripadajo razredom iz prve ali druge učne množice. Preverili bomo, ali zamrznitev vseh plasti razen zadnje odpravi oz. omili katastrofalno pozabljanje. V drugi fazi učenja ne bomo ohranili nobenih podatkov iz prve učne množice.

Shema na Sliki 3.2 prikazuje stanje naše mreže v drugi fazi učenja, če mrežo ohranimo popolnoma isto, shema na Sliki 3.3 pa stanje, če zamrznemo začetni del mreže. Za učenje v drugi fazi bomo uporabili dve optimizacijski metodi ter primerjali hitrost in obseg katastrofalnega pozabljanja, skupaj bomo izvedli 4 ponovitve eksperimenta.



Slika 3.2: Shema stanja nespremenjene mreže pri naivnem pristopu. Zelena barva označuje, da se parametri delov mreže lahko spreminjajo.

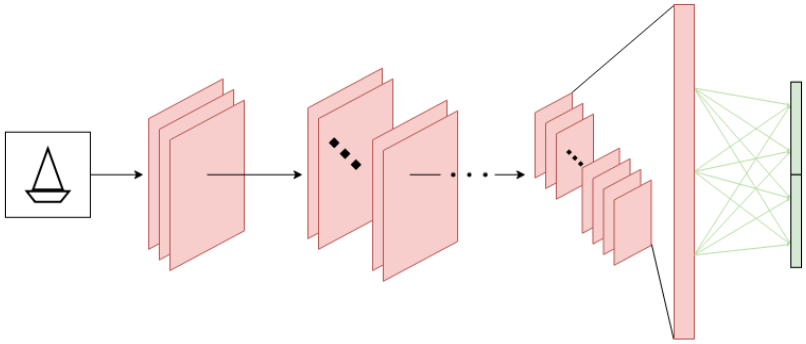
3.1.1.2 Pomnjenje podatkov iz prvotne množice

Nato bomo raziskali, kaj se dogaja, če mrežo pustimo popolnoma odmrznjeno kot na shemi na Sliki 3.2, vendar v drugi fazi učenja ohranimo tudi del podatkov iz prve učne množice. Preverili bomo, kako delež ohranjenih podatkov vpliva na katastrofalno pozabljanje oz. na klasifikacijsko točnost na prvi množici. Ravno tako bomo preverili, kaj se zgodi, če namesto celotne druge učne množice uporabimo samo določen del le-te.

3.1.1.3 Zamrznitev zadnje plasti

V tretjem eksperimentu bomo videli, kaj se zgodi, če v drugi fazi učenja zamrznemo parametre nevronov v zadnji plasti, ki pripadajo razredom iz prve faze učenja. Preverili bomo, kaj se zgodi, če poleg omenjenih nevronov zamrznemo tudi preostali del mreže. Eksperiment bomo ponovili na podmnožicah iz različnih domen in tako ocenili, kako pomembno je, če sta podmnožici iz iste domene. Preverili bomo tudi, kaj se zgodi, če sta podmnožici različnih kompleksnosti, in kako vrstni red podmnožic vpliva na končno klasifikacijsko točnost.

Shema na Sliki 3.4 prikazuje zamrznitev parametrov nevronov iz prve učne množice v zadnji plasti, shema na Sliki 3.5 pa dodatno zamrznitev



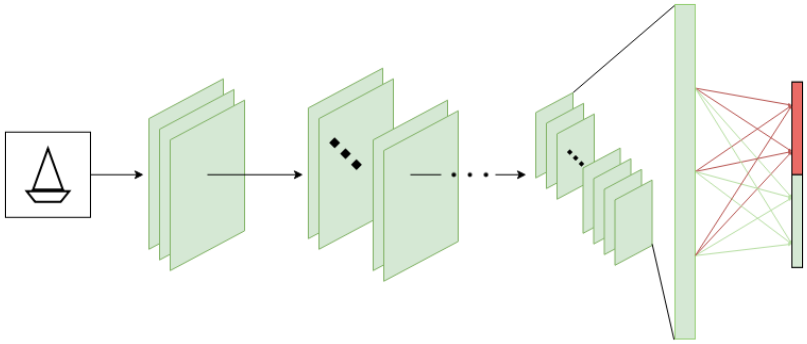
Slika 3.3: Shema stanja mreže v prvem eksperimentu, ko zamrznemo vse razen zadnje plasti. Zelena barva označuje, da se parametri delov mreže lahko spreminjajo, rdeča pa, da so zamrznjeni.

preostalega dela mreže. Tako učenje celotne mreže poteka le na delu nevronov v zadnji plasti.

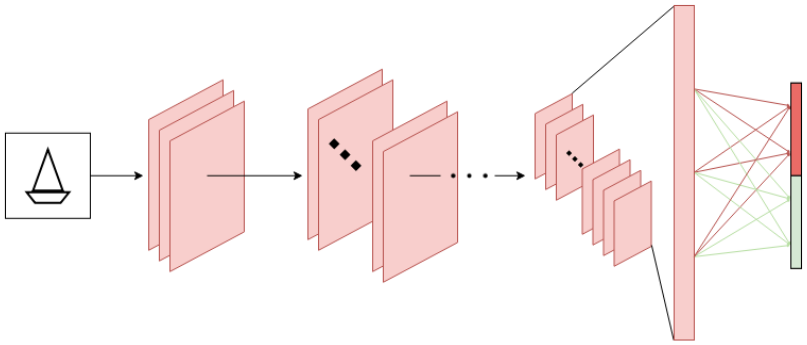
Videli bomo, da vseeno pride do katastrofalnega pozabljanja, če zamrznemo samo del nevronov iz zadnje plasti, preostale mreže pa ne. Alternativno, če zamrznemo tudi preostali del mreže se katastrofalno pozabljanje ne pojavi v večji meri, vendar če imamo podmnožici različnih kompleksnosti in se najprej učimo na lažji, potem v drugi fazi učenja na kompleksnejši ne dosežemo dovolj visoke klasifikacijske točnosti in je posledično tudi skupna nižja.

3.1.1.4 Variabilna stopnja učenja

Zaradi spoznanj iz prejšnjega eksperimenta bomo preverili, kaj se zgodi, če združimo oba pristopa. Ta eksperiment bo zato predstavljal kombinacijo obeh delov iz prejšnjega. Zamrznili bomo parametre dela nevronov iz zadnje plasti, za preostale plasti pa bomo uporabili nižjo stopnjo učenja. Shema na Sliki 3.6 prikazuje, kateri parametri v četrtem eksperimentu so zamrznjeni, za katere velja normalna stopnja učenja in za katere znižana.



Slika 3.4: Shema stanja mreže ob zamrznitvi dela zadnje plasti.

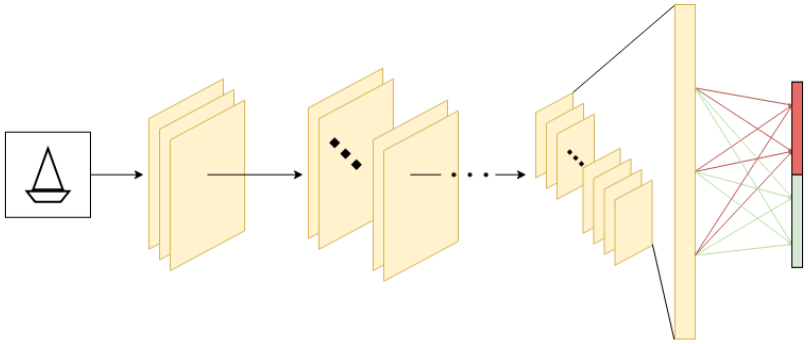


Slika 3.5: Shema stanja mreže ob zamrznitvi dela zadnje plasti in vseh preostalih plasti.

3.1.2 Orakelj

Vsi zgoraj opisani eksperimenti predpostavljajo, da imamo mreže s konstantnim številom izhodnih nevronov oz. da se arhitektura mreže ob učenju na novih razredih ne spreminja. Prav tako ob testiranju oz. uporabi mreže za posamezen testni primer ne vemo, ali pripada razredu iz prve učne podmnožice ali iz katere druge.

Alternativna verzija problema inkrementalnega učenja predpostavlja obstoj oraklja, tj. da ob uporabi nevronske mreže za posamezen primer vemo, ali ga moramo uvrstiti v enega izmed razredov, ki smo se jih učili v prvi fazi



Slika 3.6: Shema stanja mreže ob variabilni stopnji učenja. Rumena barva označuje zmanjšano stopnjo učenja, rdeča pa zamrznitev.

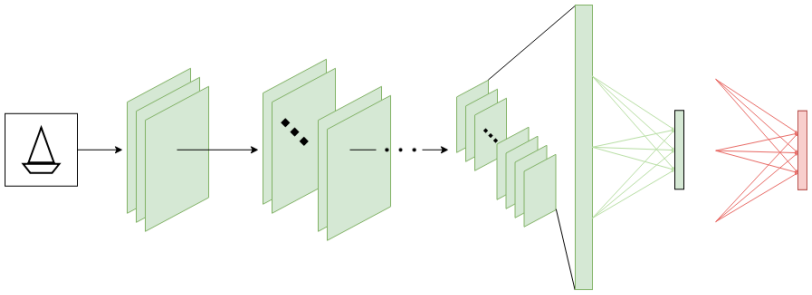
učenja, ali v enega izmed razredov iz ostalih faz učenja. Problem tako postane veliko lažji, saj primera iz ene faze učenja ne moremo napačno uvrstiti v razred iz druge faze učenja. Ravno to, da je primer uvrščen v razred iz napačne podmnožice, je namreč eden izmed glavnih dejavnikov pri katastrofalnem pozabljanju.

Ob uporabi oraklja se spremeni tudi arhitektura nevronske mreže. Namesto da imamo samo eno zadnjo plast s toliko nevroni, kot je vseh razredov iz vseh učnih podmnožic, se za vsako učno podmnožico zamenja zadnjo plast z novo, ki ima toliko izhodnih nevronov, kot je razredov v tisti učni podmnožici. Staro plast (oz. vse parametre le-te) je potrebno shraniti, saj se bo uporabljala ob testiranju. Ob testiranju mreže moramo zadnjo plast ustrezno nastaviti glede na to, kateri učni podmnožici pripada testni primerek.

3.1.2.1 Pristop z orakljem

Preverili bomo, kako izrazito je katastrofalno pozabljanje ob uporabi oraklja. Na začetku bo mreža imela zadnjo plast s toliko izhodnimi nevroni, kot je razredov v prvi učni množici. Ob začetku faze bomo odstranili in shranili naučeno zadnjo plast, na mrežo pa dodali novo s toliko nevroni, kot je razredov v drugi učni množici, in jo naučili na drugi učni množici. Ob testiranju

bomo najprej testirali na drugi testni množici, nato zamenjali zadnjo plast s staro in testirali še na prvi množici.



Slika 3.7: Shema mreže ob uporabi oraklja

Shema na Sliki 3.7 prikazuje dogajanje ob uporabi oraklja v testni fazi. Ko določeno plast odstranimo in jo shranimo, se njeni parametri ne spreminjajo več. Ostali del mreže je nespremenjen, noben parameter ni zamrznjen.

3.1.3 MAS

Memory Aware Synapses [1] je eden izmed pristopov k odpravljanju katastrofalnega pozabljanja pri inkrementalnem učenju z orakljem. Temelji na regularizaciji, in sicer kaznuje (zmanjšuje) spremembe parametrov, ki imajo močnejši vpliv na izhod iz nevronske mreže. Zagotavlja, da so spremembe parametrov, ki so pomembni za prejšnjo nalogo, omejene, medtem ko se lahko manj pomembni parametri bolj prilagodijo, da mreža doseže večjo klasifikacijsko točnost na novih nalogah. Za vsak parameter θ_i se izračuna občutljivost izhoda mreže na le-ta parameter Ω_i, tj. parcialni odvod izhodov mreže glede na parameter θ_i

$$\Omega_i = \frac{1}{N} \sum_{k=1}^{N} \left\| \frac{\partial(F(x_k))}{\partial \theta_i} \right\|, \tag{3.1}$$

kjer $F(x_k)$ predstavlja izhod iz nevronske mreže ob vhodu x_k, N pa število vseh primerov, na katerih računamo občutljivost. Občutljivost izračunamo na množici primerov, lahko je to učna množica ali del te, lahko je tudi testna

množic. Dodatna prednost te metode je, da je množica lahko neoznačena. Ker bi v (3.1) morali za mreže z več izhodnimi nevroni izračunati gradient za vsak nevron posebej, nadomestimo parcialni odvod izhodov mreže s parcialnim odvodom vsote kvadratov izhodov iz mreže (angl. l_2 norm)

$$\Omega_i = \frac{1}{N} \sum_{k=1}^{N} \left\| \frac{\partial(l_2(F(x_k)))}{\partial \theta_i} \right\| \tag{3.2}$$

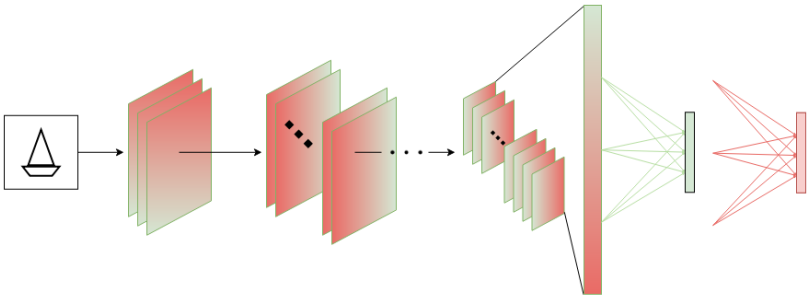
Ko imamo izračunane vse θ_i lahko prilagodimo kriterijsko funkcijo mreže, da upošteva spremembe parametrov

$$L(\Theta) = L_n(\Theta) + \lambda \sum_{\forall i} \Omega_i (\theta_i - \theta_i^*)^2, \tag{3.3}$$

kjer je $L_n(\Theta)$ prvotna vrednost kriterijske funkcije, λ predstavlja hiperparameter, ki določa stopnjo regularizacije, $\theta_i - \theta_i^*$ pa razliko med začetno in trenutno vrednostjo parametra θ_i.

3.1.3.1 Metoda MAS

Eksperimentalno bomo preverili, kako uporaba metode MAS zmanjšuje katastrofalno pozabljanje in kakšen vpliv ima vrednost parametra λ na klasifikacijsko točnost.

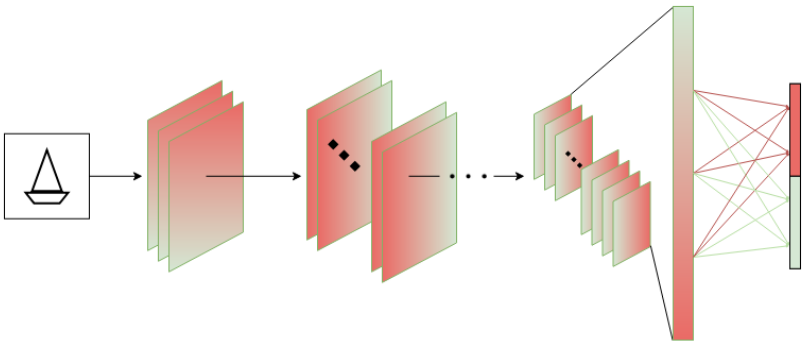


Slika 3.8: Shema mreže ob uporabi oraklja in metode MAS za odpravljanje katastrofalnega pozabljanja.

Shema na Sliki 3.8 prikazuje dogajanje ob učenju z orakljem in uporabi metode MAS. Zeleno-rdeče komponente v shemi prikazujejo, da se posamezni parametri spreminjajo skladno z njihovo vrednostjo Ω.

3.1.3.2 MAS brez oraklja

Preverili bomo, kako se principi zajeti v metodi MAS obnesejo, če oraklja nimamo na voljo. Metodo bomo prilagodili, da deluje tudi na mrežah s samo eno izhodno plastjo. Poleg tega bomo zamrznili del zadnje plasti. Shema na



Slika 3.9: Shema mreže ob uporabi metode MAS na mreži z eno izhodno plastjo.

Sliki 3.9 prikazuje stanje ob uporabi metode MAS na mreži z eno samo izhodno plastjo. Parametri zadnje plasti nevronov, ki pripadajo prej naučenim razredom, se zamrznejo, parametri zadnje plasti nevronov, ki pripadajo razredom iz trenutne učne množice, pa se prosto spreminjajo. Parametri ostalih plasti se spreminjajo glede na njihove vrednosti Ω.

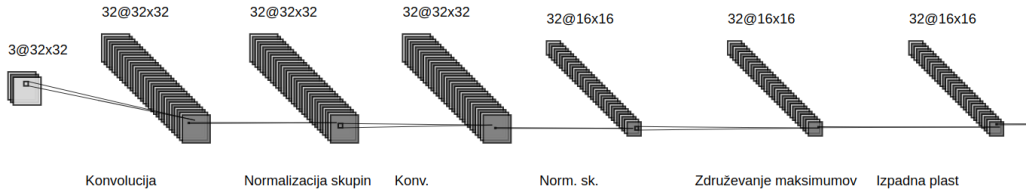
3.2 Arhitektura eksperimentalne mreže

Pristopi, opisani v Razdelku 3.1, so splošni in se lahko uporabijo na poljubnih arhitekturah nevronskih mrež. Izvedli bomo relativno veliko eksperimentov,

poleg tega bomo za vsakega poskušali najti relativno dobre vrednosti hiperparametrov, zato si za izvajanje eksperimentov ne moremo privoščiti uporabe katere izmed trenutno najboljših nevronskih mrež za klasifikacijo, kot sta recimo ResNet [8] ali DenseNet [10], saj učenje le teh na strojni opremi, ki nam je na voljo, vzame preveč časa.

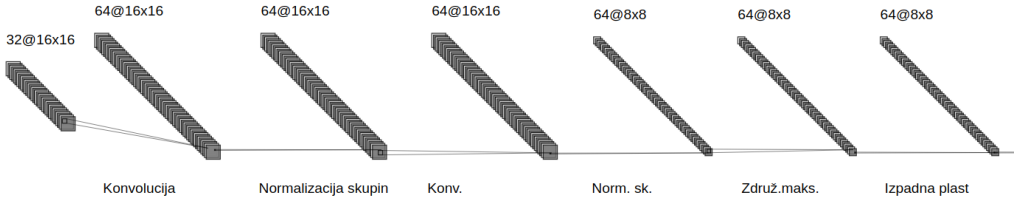
Za izvedbo opisanih eksperimentov smo zato zasnovali globoko konvolucijsko nevronsko mrežo, ki je prilagojena izbranim podatkovnim zbirkam, prav tako smo bili do določene mere omejeni pri izbiri globine oz. števila parametrov, saj lahko učenje hitro postane zelo dolgotrajno. Naša nevronska mreža sicer ne dosega tako dobrih rezultatov, kot trenutno najboljše, vendar to ne predstavlja večje ovire, saj analiziramo ozadje dogodkov, ki pripeljejo do katastrofalnega pozabljanja, poleg tega pa lahko domnevamo, da so naše ugotovitve splošne in se prenesejo tudi na večje nevronske mreže.

Zasnovana mreža je sestavljena iz $3 \times 8 + 1$ plasti. Osnovni gradnik je sestavljen iz konvolucije + ELU + normalizacije skupin + konvolucije + ELU + normalizacije skupin + združevanja z maksimizacijo + izpadne plasti. Osnovni gradnik se ponovi trikrat, na koncu je dodana polno povezana plast. Vhodne dimenzije mreže znašajo $32 \times 32 \times 3$, za vhodne dimenzije 32×32 smo se odločili zaradi uporabe podatkovnih zbirk, ki ne presegajo te velikosti.

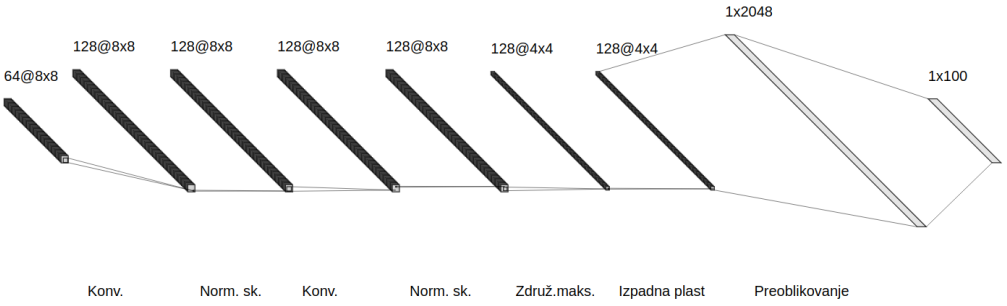


Slika 3.10: Prvi osnovni gradnik, vhod je paket slik

Edini gradnik mreže, ki se med eksperimenti lahko spreminja, je zadnja, polno povezana plast, saj imajo lahko različni eksperimenti različno število razredov (izhodnih nevronov).



Slika 3.11: Drugi osnovni gradnik, vhod je izhod iz prvega



Slika 3.12: Tretji osnovni gradnik in polno povezana plast, vhod je izhod iz drugega.

Na Slikah 3.10, 3.11 in 3.12 je predstavljena celotna eksperimentalna mreža. Ob vsaki konvoluciji se izvede tudi transformacija z aktivacijsko funkcijo ELU z vrednostjo $\alpha = 1$.

Poglavje 4

Eksperimenti

4.1 Podatkovne zbirke

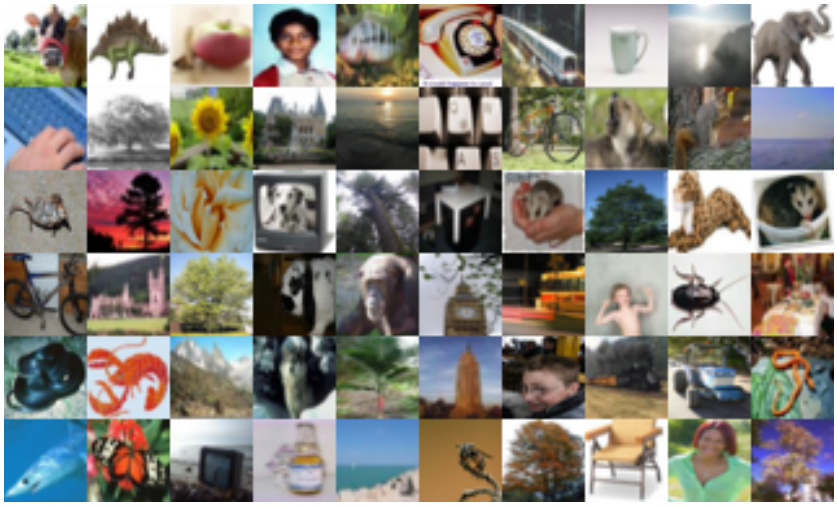
Pri izbiri podatkovnih zbirk za izvedbo eksperimentov smo bili delno omejeni tudi z razpoložljivo strojno opremo, zato smo se odločili za relativno manjše podatkovne zbirke, tako po velikosti primerov kot po številu le-teh. Izbrane podatkovne zbirke so v raziskovalnih delih zelo pogoste, zato lahko sklepamo, da so kljub temu ustrezne za ovrednotenje naših ugotovitev, le-te pa bi se obdržale tudi ob uporabi večjih podatkovnih zbirk.

4.1.1 CIFAR-100

CIFAR-100 [14] je zbirka 60.000 barvnih fotografij velikosti 32×32, razdeljenih v 100 razredov. 50.000 fotografij predstavlja učno množico, ostalih 10.000 pa testno. Vsaka izmed kategorij ima 500 učnih in 100 testnih slik. Na fotografijah so jasno vidni objekti, postavljeni na sredino slike. Zaradi razmeroma velikega števila učnih primerov in majhnih dimenzij je to ena izmed podatkovnih zbirk, ki se zelo pogosto uporablja za eksperimente, povezane s klasifikacijo.

4.1.2 CIFAR-10

CIFAR-10 [14] je sestavljena iz istih fotografij kot CIFAR-100, le da so le-te razdeljene v 10 razredov namesto 100. Vse ostale lastnosti so popolnoma identične. Tudi ta zbirka je iz enakih razlogov zelo pogosto uporabljena.



Slika 4.1: Izsek fotografij iz podatkovne zbirke CIFAR-100.

4.1.3 Fashion-MNIST

Fashion-MNIST [21] je zbirka 70.000 črno-belih fotografij oblačil in modnih dodatkov. Zbirka je bolj kompleksna kot MNIST [15] in je bila zasnovana z namenom, da le-to nadomesti kot enega izmed standardov za evalvacijo modelov. Sestavljena je iz 70.000 črno-belih fotografij velikosti 28×28, razdeljenih v 10 razredov, s 6.000 učnimi in 1.000 testnimi primeri za vsak razred. Objekti so centrirani, lepo vidni in so edina stvar na slikah. Tako kot CIFAR-100 je tudi ta zbirka popularna zaradi svoje kompaktnosti. V eksperimentih, kjer bomo uporabili to podatkovno zbirko, bomo fotografije povečali na dimenzije 32×32 in pretvorili v RGB barvni prostor.

Slika 4.2: Izsek fotografij iz podatkovne zbirke Fashion-MNIST.

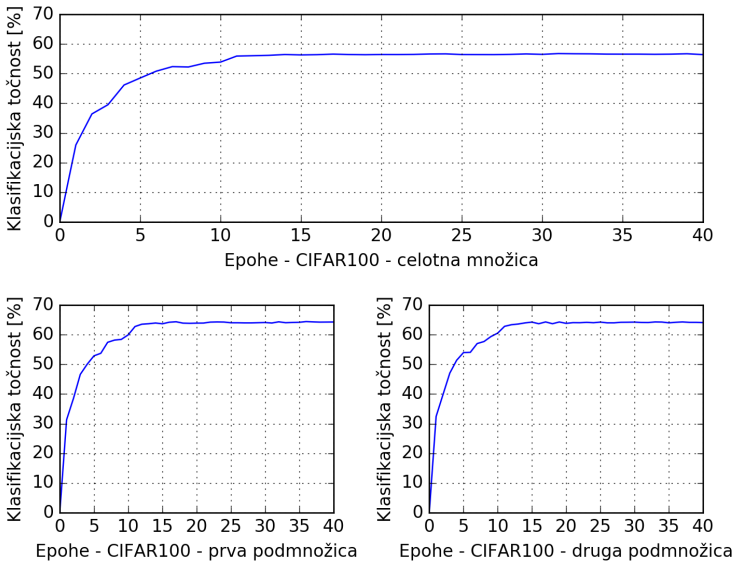
4.2 Eksperimenti

Idealna rešitev problema katastrofalnega pozabljanja bi omogočala, da bi klasifikacijska točnost ob učenju na dveh ločenih množicah podatkov v dveh fazah bila enaka oz. vsaj približno enaka, kot ob učenju na združeni množici. Za pristop z orakljem pa bi rešitev pomenila, da se klasifikacijska točnost na starih podatkih ne bi opazno zmanjšalo ob učenju na novih podatkih.

Grafi na Sliki 4.3 in Sliki 4.4 prikazujejo klasifikacijske točnosti, ki jih dobimo ob učenju na posameznih podmnožicah in združeni množici. Želeli bi se čim bolj približati klasifikacijskim točnostim, ki jih dobimo ob učenju na združenih množici.

4.2.1 Klasični scenarij

Začeli bomo z izvedbo eksperimentov, v katerih ob uporabi mreže ne moremo omejiti množice razredov, v katere lahko klasificiramo primer, oraklja torej nimamo na voljo.

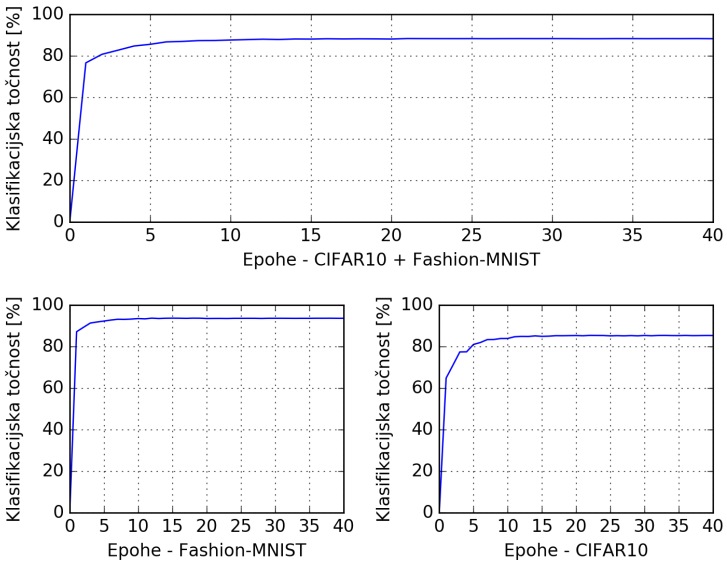


Slika 4.3: Klasifikacijske točnosti pri učenju na celotni zbirki CIFAR100 (zgoraj) in pri učenju na dveh podmnožicah (spodaj).

4.2.1.1 Naivni pristop

Nevronsko mrežo učimo na dveh podmnožicah podatkovne zbirke CIFAR-100, vsaka vsebuje 50 razredov in vse učne in testne primere, ki pripadajo tem razredom. Mreža ima že na začetku 100 izhodnih nevronov, lahko bi sicer začeli samo s 50 in nato ob učenju druge množice dodali še preostalih 50, vendar se za to nismo odločili. Mrežo najprej naučimo na prvih 50 razredih, učimo za 40 epoh, z začetno stopnjo učenja 0,001, ki se zmanjša za faktor 0,1 vsakih 10 epoh, velikost paketa je 64. Optimizacijska metoda je Adam, kriterijska funkcija je križna entropija, dosežemo približno 63% klasifikacijsko točnost na testni množici.

Nato isto mrežo učimo na preostalih 50 razredih, med učnimi podatki ni nobenega izmed prvotnih 50 razredov. Za optimizacijsko metodo uporabimo SGD in Adam, za obe preverimo tudi, kaj se zgodi, če zamrznemo preostali del mreže. Shemi sta prikazani na Sliki 3.2 in Sliki 3.3, v kolikor zamrznemo



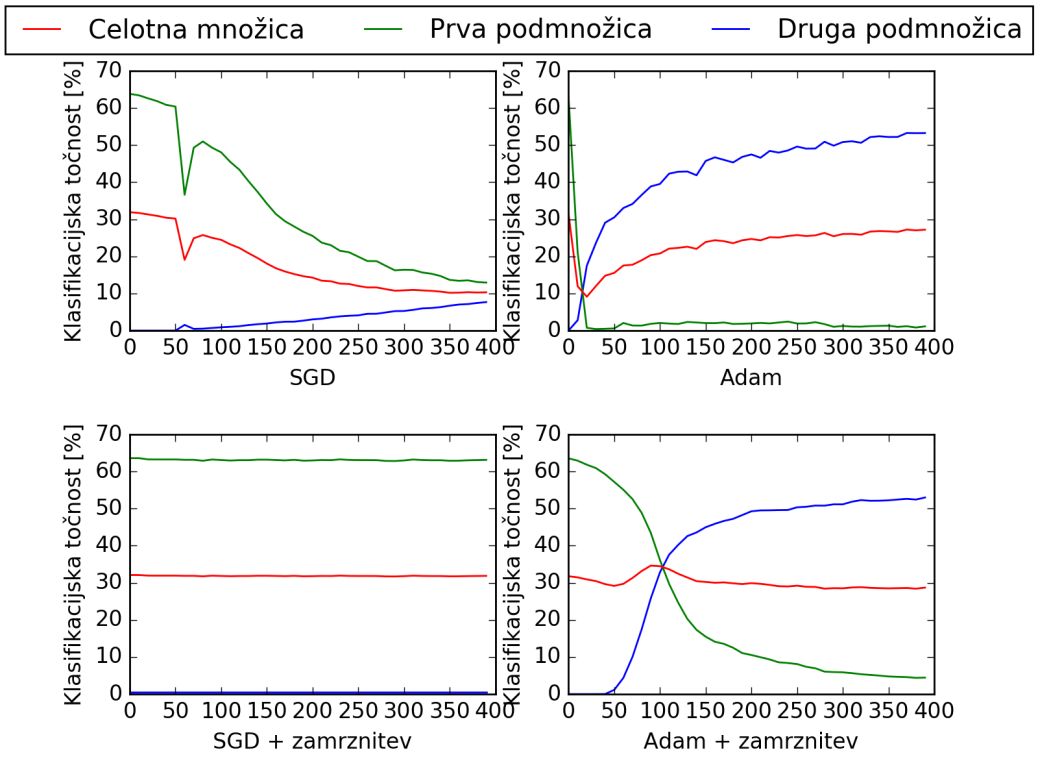
Slika 4.4: Klasifikacijske točnosti pri učenju na zbirkah Fashion-MNIST (levo spodaj), CIFAR10 (desno spodaj) in združeni zbirki (zgoraj).

preostali del mreže. Ostali parametri učenja so identični prvotnim. Skupno izvedemo 4 ponovitve eksperimenta.

Graf na Sliki 4.5 prikazuje spreminjanje klasifikacijske točnosti obeh testnih množic v prvih 80-ih iteracijah prve epohe učenja na drugi testni množici. Vidimo, da mreža brez zamrznitve nemudoma pozabi praktično vse znanje o prvi množici, ne glede na to, katero optimizacijsko metodo uporabimo, čeprav je pozabljanje počasnejše pri uporabi SGD.

V kolikor zamrznemo preostali del mreže, se katastrofalno pozabljanje upočasni, vendar je ob uporabi Adam-a še vedno močno. Edini primer, v katerem mreža po eni epohi učenja ohrani del znanja o prvi množici podatkov, je, ko zamrznemo preostali del mreže in za optimizacijsko metodo uporabimo SGD.

Graf na Sliki 4.6 prikazuje spreminjanje klasifikacijske točnosti po epohah, če zamrznemo mrežo in za optimizacijsko metodo uporabimo SGD. Katastro-

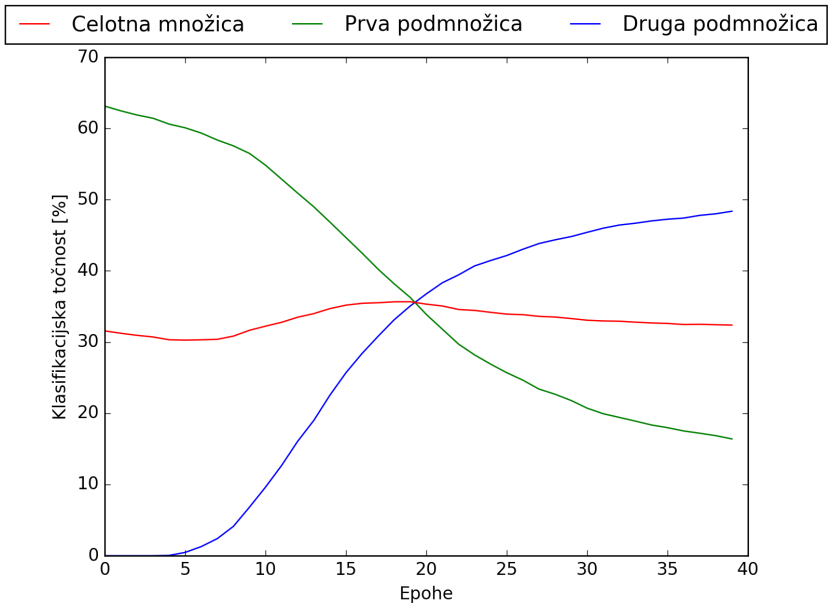


Slika 4.5: Katastrofalno pozabljanje v prvi epohi učenja na drugi učni množici.

falno pozabljanje se sicer močno upočasni, vendar mreža še vedno pozabi vse zakonitosti prve učne množice.

K tako izrazitemu in hitremu pozabljanja pripomore več faktorjev. Eden izmed glavnih je verjetno popolna odsotnost učnih primerov iz prve učne množice v drugi fazi učenja. Neuravnotežena sestava učne množice je dokaj dobro raziskan problem [2], obstaja tudi nekaj načinov za odpravljanje letega, vendar v našem primeru noben ne pride v upoštev, saj so podatki iz prve učne množice popolnoma odsotni.

Domnevamo, da se mreža ob odsotnosti primerov iz prve učne množice preprosto nauči, da ne sme nobenega testnega primera prepoznati, kot da ta pripada razredu iz prve faze učenja. Za boljši vpogled bomo opazovali



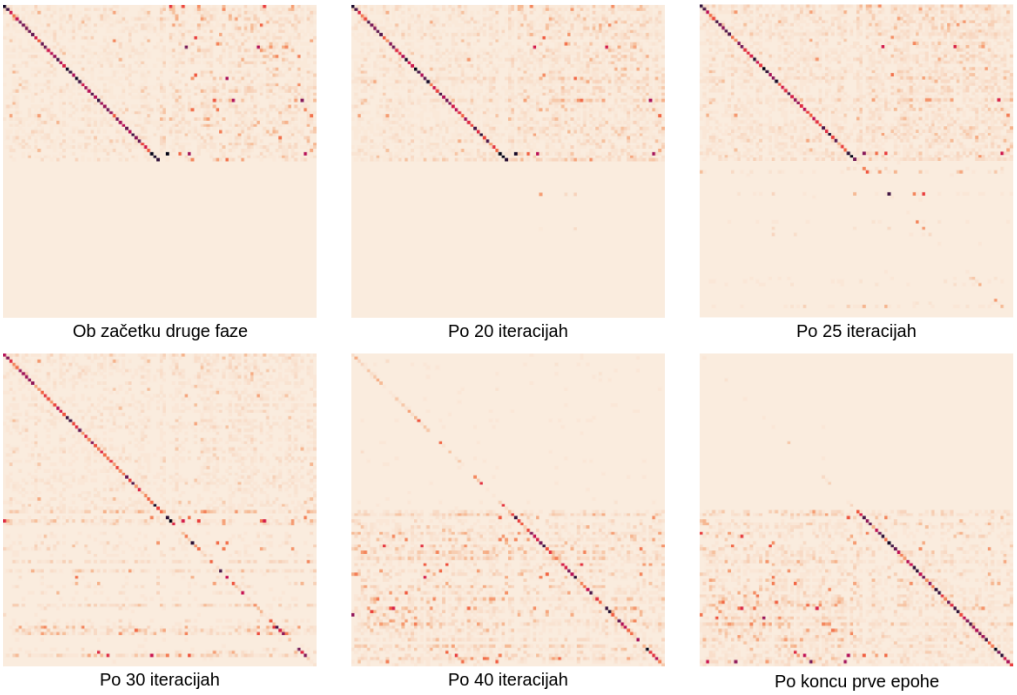
Slika 4.6: Upočasnjeno katastrofalno pozabljanje ob uporabi SGD in zamrznitvi mreže.

spreminjanje matrik zamenjav v drugi fazi učenja, ob zamrznitvi mreže in uporabi Adam-a (spodnji desni graf na Sliki 4.5).

Matrike zamenjav na Sliki 4.7 delno potrjujejo našo domnevo, da je eden izmed glavnih problemov neuravnotežena učna množica. Stolpec predstavlja resnični razred, vrstica pa napovedan. Opazimo, da v prvih iteracijah mreža prepozna vse primere, kot da pripadajo razredom iz prvotne učne množice, nato pa razmeroma hitro praktično vse primere prepozna, kot da pripadajo razredom iz druge učne množice.

Da bi bolje razumeli, zakaj pride do takega padca, smo raziskali, kaj se dogaja s parametri mreže. Preverili smo, kako se spreminjajo uteži in odmiki v zadnji, polno povezani plasti mreže.

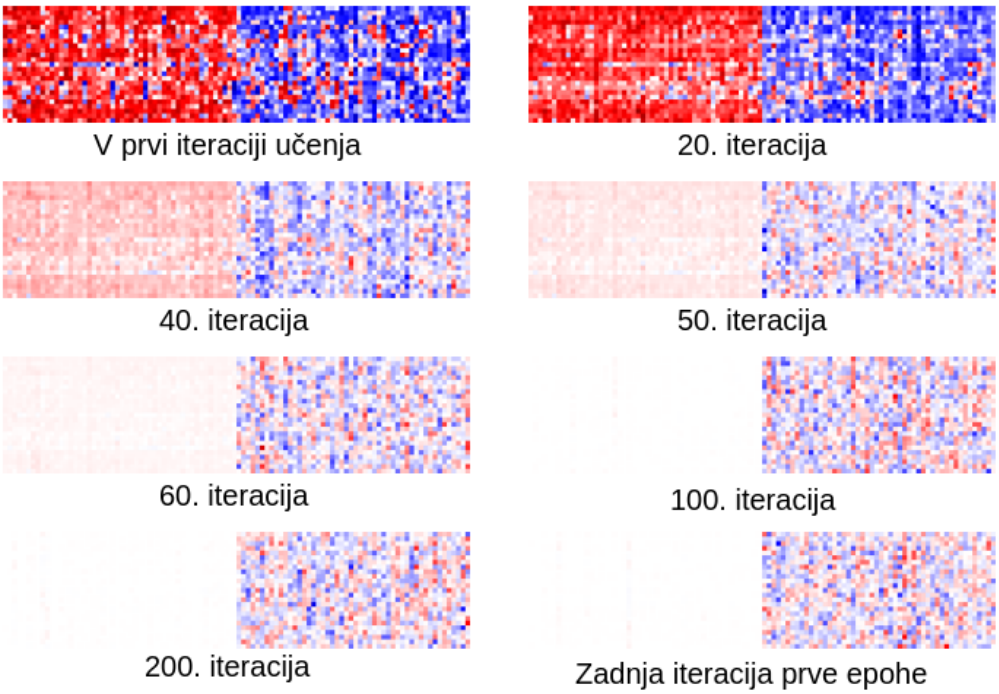
Na Sliki 4.8 lahko spremljamo, kako se po iteracijah prve epohe druge faze učenja spreminjajo uteži zadnje plasti. Prikazana je razlika med vrednostimi uteži v dveh zaporednih iteracijah. Rdeča barva označuje negativno spremembo (znižanje) vrednosti, modra pa pozitivno spremembo, enaka intenzi-



Slika 4.7: Spreminjanje matrik zamenjav. Stolpec predstavlja resničen razred, vrstica pa napovedan.

teta predstavlja enako spremembo na vseh slikah. Intenziteta barve označuje relativno velikost spremembe, večja intenziteta predstavlja večjo spremembo. Posamezna vrstica v vsaki sliki predstavlja vse uteži, ki utežujejo povezave od enega nevrona v predzadnji plasti do vsakega izmed nevronov v izhodni plasti. Posamezen stolpec predstavlja vse uteži, ki gredo iz vsakega nevrona v predzadnji plasti do določenega nevrona v izhodni plasti. Celotna slika ima toliko vrstic, kot je nevronov v predzadnji plasti (v našem primeru 2048), vendar je na slikah zaradi preglednosti prikazanih le prvih 25, saj se ostale uteži obnašajo podobno kot prikazane.

Takoj opazimo, da se uteži, ki pripadajo nevronom, ki predstavljajo razrede iz prve učne množice, spreminjajo drugače kot tiste, ki pripadajo nevronom, ki predstavljajo razrede iz druge učne množice. V začetnih iteracijah

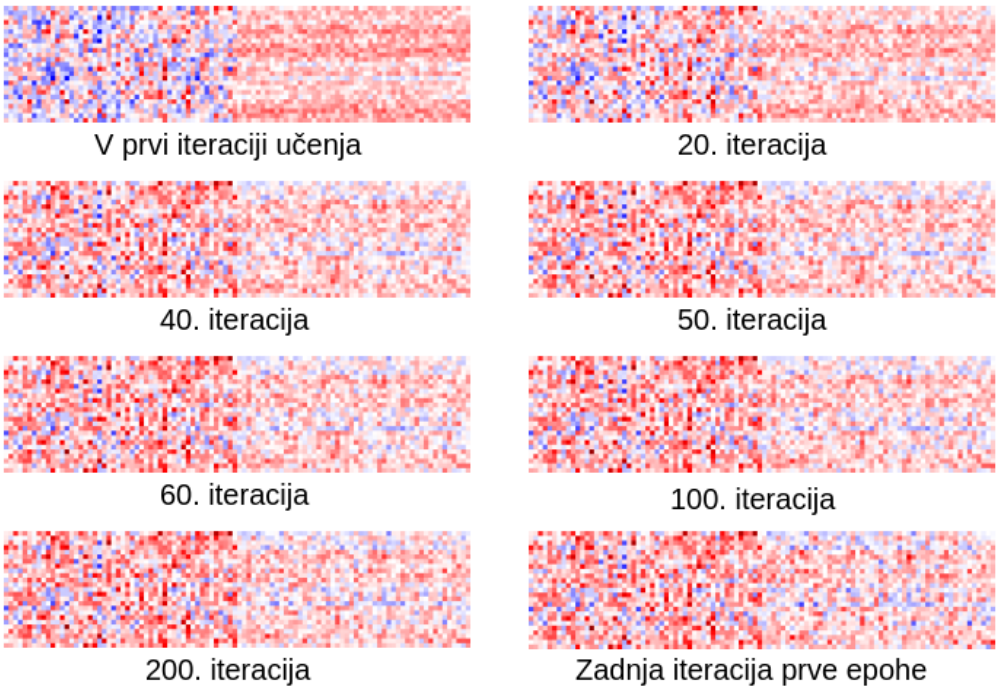


Slika 4.8: Spreminjanje uteži od prvih 25-ih nevronov predzadnje plasti do nevronov v izhodni plasti.

je opazen izrazit trend manjšanja uteži nevronov, ki predstavljajo razrede iz prve učne množice, in hkratnega višanja uteži nevronov, ki predstavljajo razrede iz druge učne množice. Sklepamo lahko, da se mreža hitro nauči, da ne sme prepoznavati testnih primerov, kot da pripadajo razredom iz prve učne množice.

Po približno 60 iteracijah učenja na drugi učni množici mreža praktično nobenega primera ne prepozna več, kot da pripada razredom iz prve učne množice (glej matrike zamenjav na Sliki 4.7 in spodnji desni graf na Sliki 4.5). Zgoraj opažen trend zniževanja vrednosti uteži nevronov, ki predstavljajo razrede iz prve učne množice, se zato ustavi, uteži nevronov, ki pripadajo razredom iz druge učne množice, pa se poljubno spreminjajo.

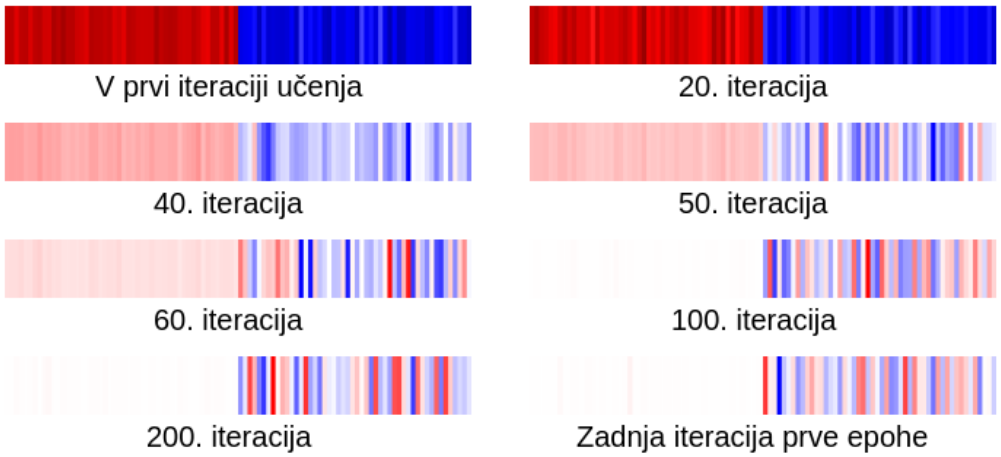
Slika 4.9 prikazujejo dejanske vrednosti uteži zadnje plasti v prvi epohi



Slika 4.9: Dejanske vrednosti uteži od prvih 25-ih nevronov predzadnje plasti do nevronov v izhodni plasti.

druge faze učenja, barve in intenzitete imajo enak pomen kot na Sliki 4.8. Opazimo lahko, da se uteži v prvi fazi učenja ob odsotnosti učnih primerov iz druge učne množice oblikujejo večinoma le za prvo polovico izhodnih nevronov. Za drugo polovico so vrednosti praviloma negativne, saj mreža še nikoli ni videla primera, ki bi jim pripadal. Ker se za optimizacijsko funkcijo uporablja Adam, se že znotraj ene epohe trend rahlo obrne, uteži do prve polovice nevronov gredo večinoma v negativne vrednosti, uteži do druge polovice nevronov pa dobijo več pozitivnih vrednosti.

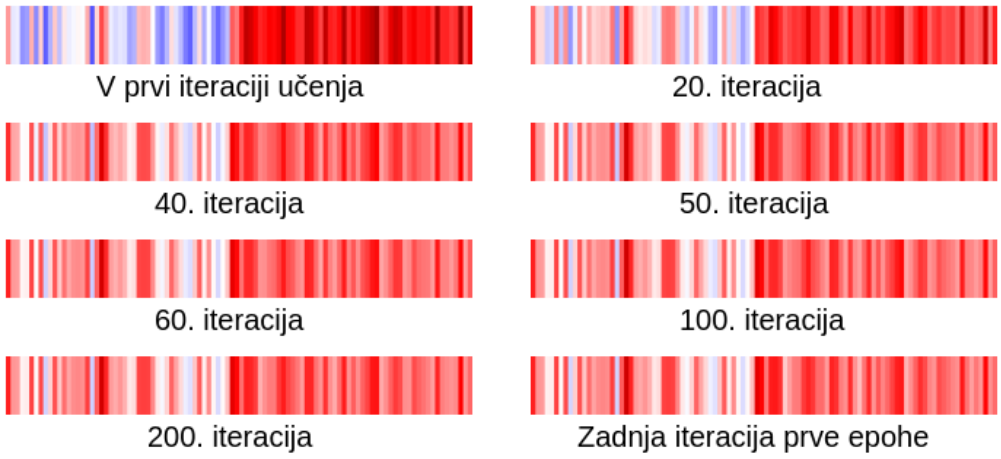
Na Sliki 4.10 so predstavljene spremembe odmikov. Barve in intenzitete imajo enak pomen kot na Sliki 4.8, posamezen stolpec pa predstavlja odmik določenega nevrona v izhodni plasti. Vidimo, da veljajo podobne zakonitosti kot za uteži.



Slika 4.10: Spreminjanje odmikov nevronov v izhodni plasti.

Slike 4.11 prikazujejo dejanske vrednosti odmikov nevronov v izhodni plasti. Barve in intezitete imajo enak pomen kot na Sliki 4.8. Ponovno lahko vidimo, da so odniki druge polovice nevronov izrazito negativni, odmiki prve polovice nevronov pa večinoma pozitivni. Že v eni epohi se odmiki prve polovice nevronov izrazito zmanjšajo, odmiki druge polovice nevronov pa postanejo manj negativni.

Preverili smo tudi, koliko se spreminjajo vsi parametri mreže ob učenju na obeh učnih množicah. Graf na Sliki 4.12 prikazuje, za koliko se razlikujejo vrednosti parametrov mreže. Prikazane so vsote absolutnih vrednosti razlik parametrov med dvema epohama. Prikazane so tako spremembe vseh parametrov mreže kot samo spremembe uteži v zadnji plasti. Vrednosti na levem grafu so normalizirane glede na padajočo stopnjo učenja, vrednosti na desnem grafu pa so dejanske spremembe. Opazimo, da so spremembe največje ob začetku učenja na novi učni množici, kar je skladno s pričakovanji. Zanimivo je, da so spremembe ob začetku druge faze učenja približno enako velike kot ob začetku prve faze, pričakovali bi namreč, da bi bile manjše.



Slika 4.11: Dejanske vrednosti odmikov nevrovnov v izhodni plasti.

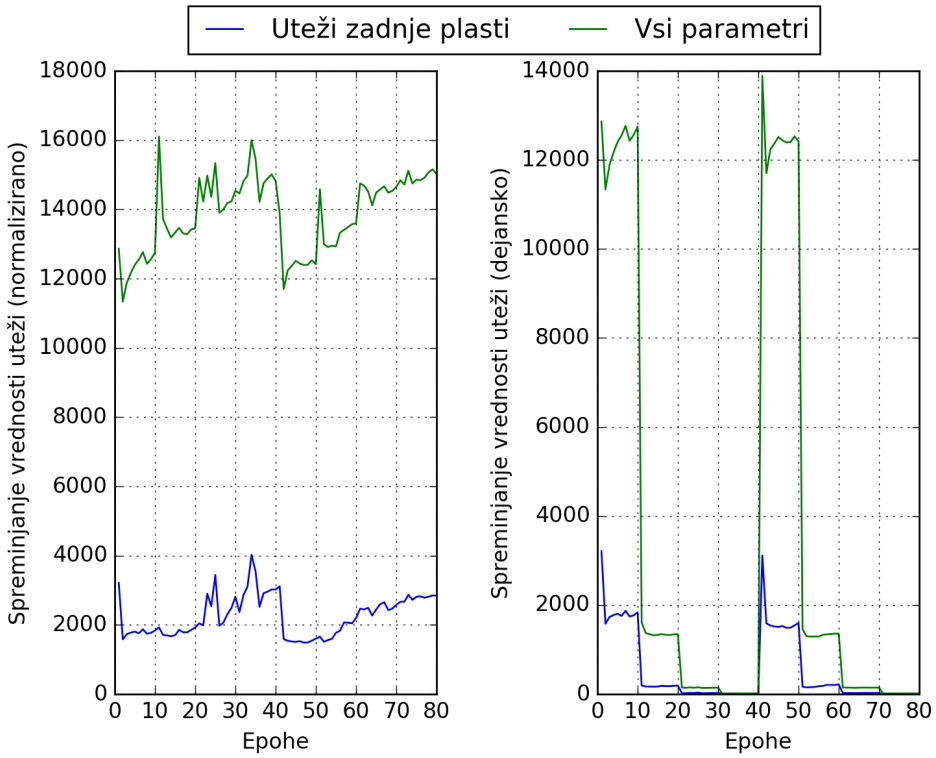
4.2.1.2 Pomnjenje podatkov iz prvotne množice

Eden izmed razlogov, da v Eksperimentu 4.2.1.1 pride do katastrofalnega pozabljanja, je odsotnost učnih primerov iz prve učne množice v drugi fazi učenja. Preverili bomo, kako vpliva delež shranjenih podatkov iz prvotne množice, ki jih pozneje uporabimo skupaj s podatki iz druge množice za učenje v drugi fazi. Parametri učenja v prvi in drugi fazi so enaki kot v Eksperimentu 4.2.1.1. Po koncu prve faze učenja zamrznemo vse plasti razen zadnje.

Shema mreže v drugi fazi učenja je prikazana na Sliki 3.3.

Eksperimenta se razlikujeta v zgradbi učne množice za drugo fazo učenja. V tem primeru ohranimo določen delež podatkov iz prvotne učne množice (glej legendo Slike 4.13). Ne glede na uporabljen delež mrežo učimo na enakem številu primerov iz obeh učnih množic, kar pomeni, da bodo posamezni shranjeni primeri iz prve množice večkrat uporabljeni znotraj iste epohe, in sicer $\frac{1}{delež}$-krat.

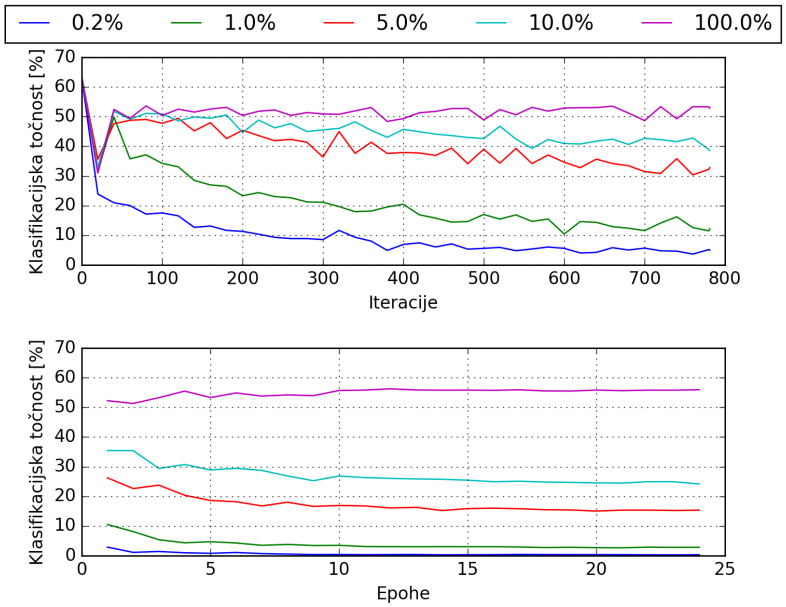
Grafa na Sliki 4.13 prikazujeta, da se s pomnjenjem primerov katastrofalno pozabljanje močno zmanjša, vendar je še vedno prisotno. Ravno tako



Slika 4.12: Spreminjanje vrednosti parametrov mreže. Na levem grafu so prikazane normalizirane vrednosti, na desnem pa dejanske spremembe.

nam Grafa na Sliki 4.14 prikažeta, da se klasifikacijska točnost na drugi testni množici zniža, če ohranimo večji del podatkov prve učne množice.

Preučili smo tudi, kako na padanje klasifikacijske točnosti vpliva delež podatkov iz druge testne množice v drugi stopnji učenja. Graf na Sliki 4.15 prikazuje, kako se klasifikacijska točnost spreminja, če za učenje v drugi fazi uporabimo 10% (leva grafa) in 50 % (desna grafa) podatkov iz druge učne množice. Opazimo, da je katastrofalno pozabljanje na prvi učni množici manjše, če ne uporabimo celotne druge učne množice. Domnevamo lahko, da mreža ohrani več znanja o prvi množici, saj ne vidi toliko primerov iz druge, vendar zaradi tega pade klasifikacijska točnost na drugi učni množici (ni prikazana na grafu).

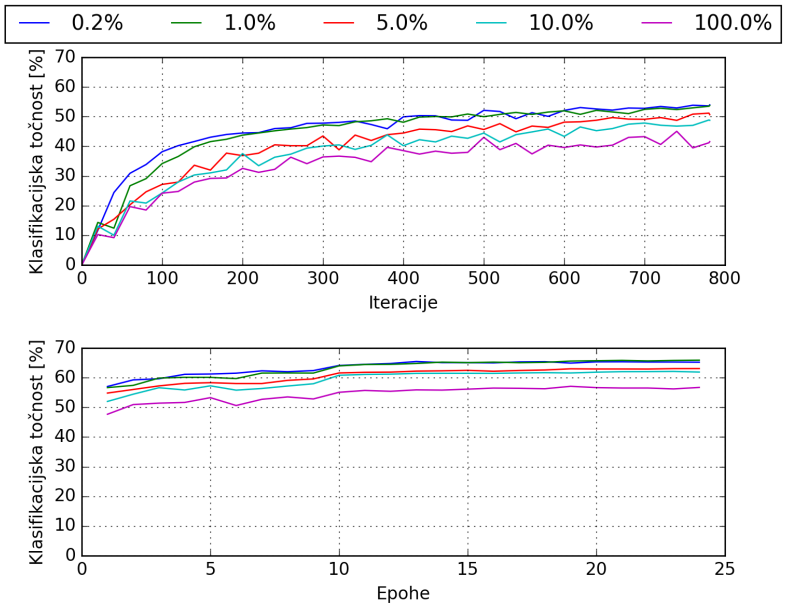


Slika 4.13: Vpliv pomnjenja primerov na pozabljanje.

4.2.1.3 Zamrznitev parametrov nevronov v izhodni plasti

Eksperiment 4.2.1.1 nazorno prikaže, da je eden izmed glavnih razlogov za katastrofalno pozabljanje izrazito spreminjanje parametrov zadnje plasti mreže, kar privede do tega, da le-ta že po eni epohi ne uvrsti skoraj nobenega testnega primera v razred iz prve učne množice. Na Slikah 4.8, 4.9, 4.10 in 4.11 vidimo, da se uteži in odmiki, ki pripadajo prvi polovici nevronov zadnje plasti, intenzivno zmanjšujejo. Zanima nas, kaj se zgodi, če zamrznemo te parametre.

Zamrznitev parametrov izvedemo tako, da nastavimo ustrezne parcialne odvode $\frac{\partial C}{\partial a^L}$ na 0 med vzvratnim razširjanjem. Tako preprečimo spremembe teh parametrov, poleg tega pa zagotovimo, da se morebitna napaka ne razširja nazaj po mreži. V drugi fazi učenja torej nastavimo vse parcialne odvode aktivacij nevronov, ki pripadajo razredom iz prve faze, na 0. Učenje zadnje plasti tako poteka le na delu nevronov, ki pripadajo razredom iz druge učne množice, kar bi moralo upočasniti katastrofalno pozabljanje.

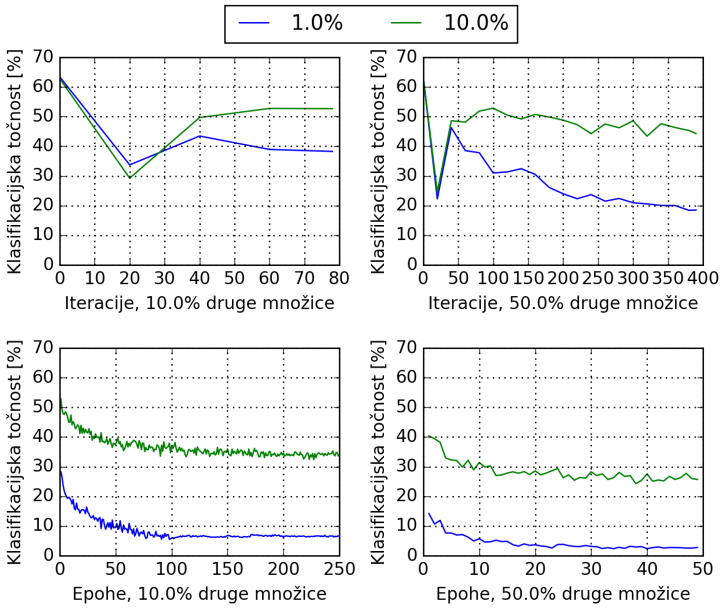


Slika 4.14: Vpliv pomnjenja primerov na doseženo klasifikacijsko točnost na drugi testni množici v drugi fazi učenja.

Slika 4.16 nazorno prikazuje, kateri parametri v zadnji plasti so zamrznjeni. Celotna shema je prikazana na Sliki 3.4.

Graf na Sliki 4.17 prikazuje, kaj se dogaja s klasifikacijskimi točnostmi na obeh in na združeni testni množici. Vidimo, da se katastrofalno pozabljanje res upočasni, vendar mreža še vedno klasificira večino primerov v razrede iz druge učne množice, saj se ostali parametri mreže preveč prilagodijo drugi učni množici. Preverili bomo, še kaj se zgodi, če poleg parametrov zadnje plasti, ki pripadajo nevronom iz prve učne množice, zamrznemo tudi preostali del nevronske mreže. Učenje celotne mreže tako poteka le na utežeh in odmikovh nevronov, ki pripadajo razredom iz druge učne množice.

Graf na Sliki 4.18 prikazuje, kaj se dogaja, če zamrznemo večino mreže. Celotna shema zamrznitve je prikazana na Sliki 3.5. Opazimo, da se katastrofalno pozabljanje zelo zmanjša, mreža kljub učenju na dveh ločenih podmnožicah doseže nekaj več kot 47% klasifikacijsko točnost na skupni testni

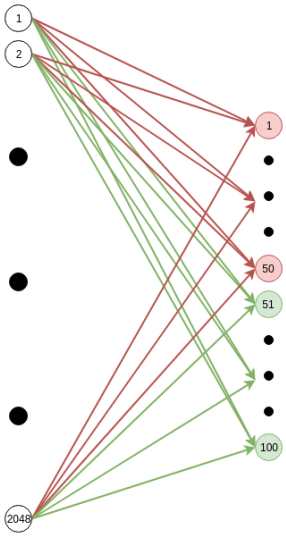


Slika 4.15: Vpliv razmerja deležev prve in druge učne množice na pozabljanje.

množici.

Matrika zamenjav na Sliki 4.19 pokaže, da mreža razmeroma enakomerno klasificira testne primere med razrede, ki pripadajo obema učnima podmnožicama.

Na Sliki 4.20 lahko vidimo, kako se izoblikujejo uteži in odmiki v zadnji plasti. Pomen posameznih elementov, barv in intenzivnosti je enak kot na Sliki 4.8, le da so pri utežeh namesto prvih 25 prikazane povezave prvih 50 nevronov. Vidimo, da se v drugi fazi izoblikujejo močnejše uteži (višja intenziteta barv) in odmiki, vendar je kljub temu klasifikacijska točnost na obeh podmnožicah približno enaka. Ena izmed možnih razlag je, da so se v učenju v prvi fazi, ko se je učila celotna mreža, izoblikovale uteži, ki bolj poudarjajo značilnosti (angl. features) slik iz prve učne množice, in so se posledično uteži v drugi fazi morale bolj prilagoditi, da so učni primeri bili ustrezno prepoznani. Kljub razliki v intenzivnosti lahko za uteži opazimo, da so pozitivne in negativne vrednosti razmeroma naključno porazdeljene, v

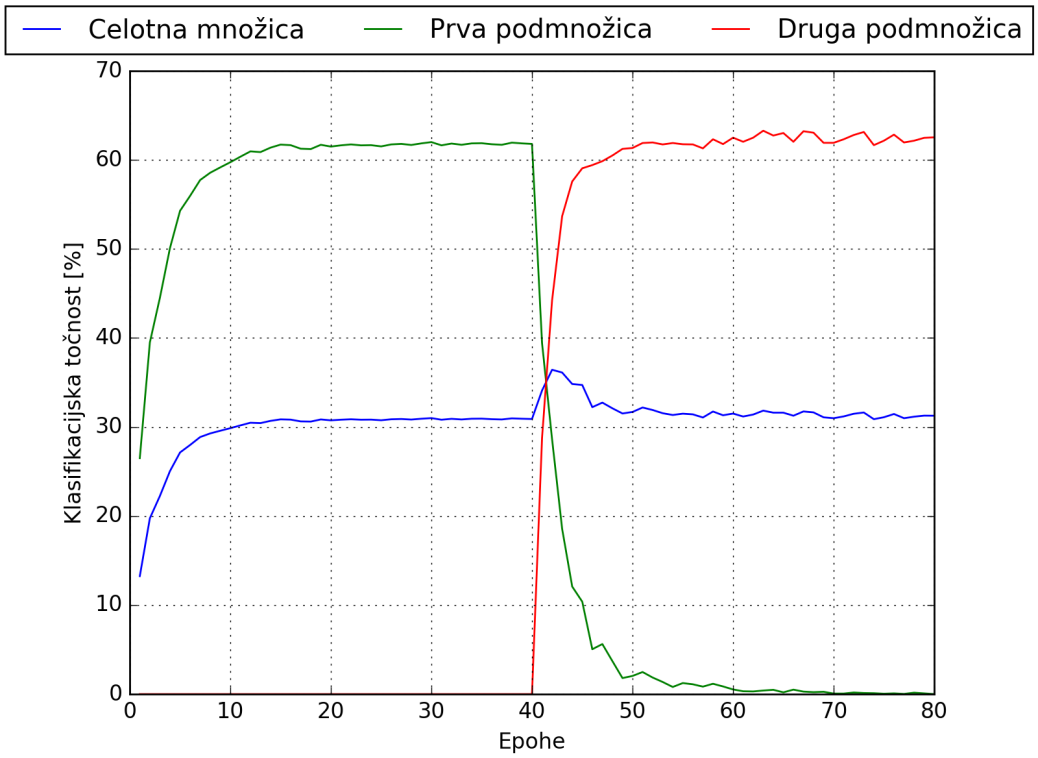


Slika 4.16: Prikaz zamrnjenih parametrov v zadnjih dveh plasteh nevronske mreže. Rdeče obarvane povezave in nevroni označujejo, da je utež oz. odmik zamrznjena, zelena označuje, da ni.

čemer se močno razlikujejo od tistih na Sliki 4.9.

Rezultati, ki jih dobimo z zamrznitvijo večine mreže, so sicer spodbudni, vendar se poraja vprašanje, ali so tako dobri le zato, ker je domena obeh učnih podmnožic enaka; obe sta namreč sestavljeni iz polovice podatkovne zbirke CIFAR-100. Domnevo bomo preverili tako, da bomo mrežo namesto na dveh podmnožicah zbirke CIFAR-100 učili na dveh zbirkah, CIFAR-10 in Fashion-MNIST. Eno učno podmnožico torej predstavlja zbirka CIFAR-10, drugo pa Fashion-MNIST.

Grafa na Sliki 4.21 prikazujeta, kaj se dogaja, če mrežo učimo na dveh podatkovnih zbirkah, ki nista iz popolnoma enake domene. Opazimo lahko, da zamrznitev parametrov mreže katastrofalno pozabljanje močno zmanjša, kljub temu da sta podmnožici iz različnih domen. Druga lastnost, ki je lepo vidna iz grafov, je da ima vrstni red podmnožic močan vpliv na končno klasifikacijsko točnost. V kolikor najprej učimo na CIFAR-10 (zgornji graf), ki je bolj kompleksna zbirka, se mreža nauči značilk, ki so dovolj opisne, da

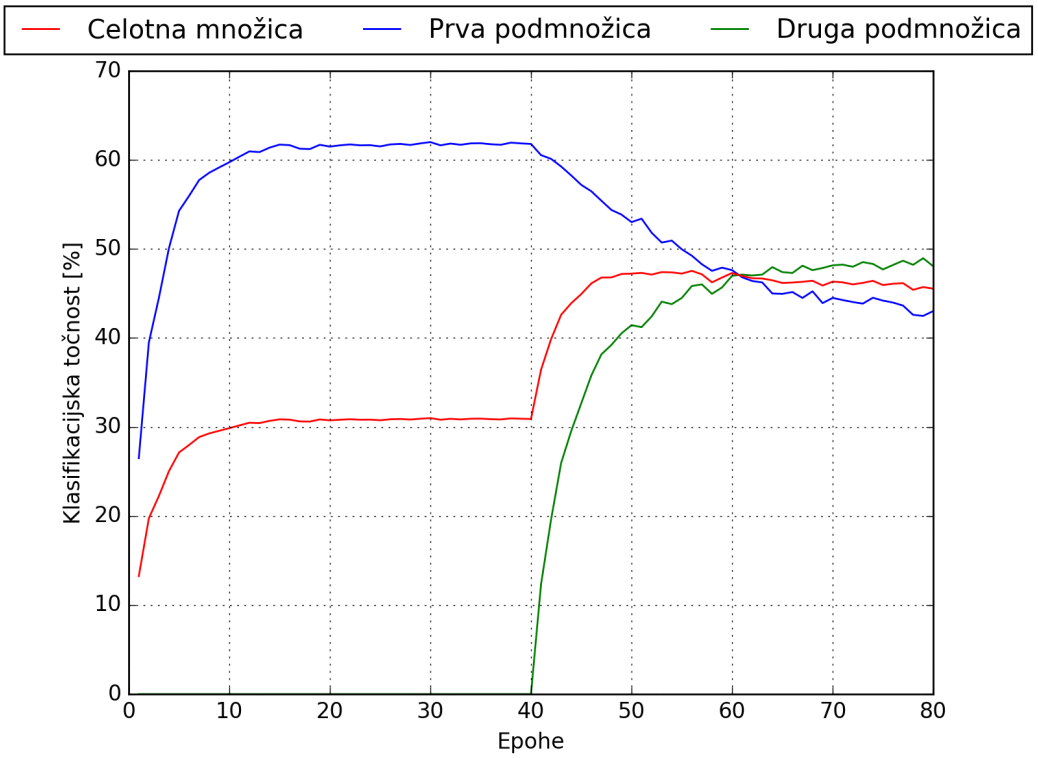


Slika 4.17: Klasifikacijske točnosti ob zamrznitvi parametrov nevronov zadnje plasti, ki pripadajo razredom iz prve učne množice.

lahko mreža z njimi dobro prepozna tudi primere iz Fashion-MNIST. Obratno ne velja, saj značilke, ki se jih mreža nauči na Fashion-MNIST, ne zadostujejo za kvalitetno prepoznavo primerov iz CIFAR-10. Vidimo, da je klasifikacijska točnost, ki jo dobimo, če najprej učimo na CIFAR-10, za približno 25% višja od tiste, ki jo dobimo, če najprej učimo na Fashion-MNIST.

4.2.1.4 Variabilna stopnja učenja

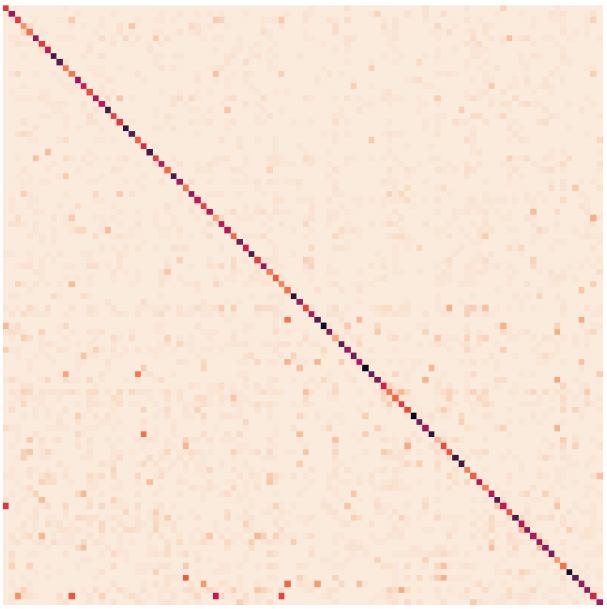
Na spodnjem grafu na Sliki 4.21 vidimo, da se mreža ne more dobro naučiti značilnosti druge podmnožice, v kolikor značilke iz prve stopnje učenja niso dovolj kvalitetne. Prav tako graf na Sliki 4.17 pokaže, da je katastrofalno pozabljanje še vedno prisotno, v kolikor ne zamrznemo vseh razen zadnje plasti



Slika 4.18: Klasifikacijske točnosti ob zamrznitvi parametrov nevronov zadnje plasti, ki pripadajo razredom iz prve učne množice, in vseh ostalih plasti mreže.

v mreži. Preverili bomo, kaj se zgodi, če sprejmemo kompromis. Mreže ne bomo zamrznili, vendar bomo za vse plasti razen zadnje uporabili zmanjšano stopnjo učenja. Zanima nas, ali bo to mreži omogočilo, da se lahko dovolj prilagodi podatkom iz druge faze učenja, ne da bi pozabila vse zakonitosti prvotnih. Zmanjšanje stopnje učenja izvedemo tako, da v mrežo pred zadnjo plast dodamo novo, ki v prehodu naprej le vrne vhod, v vzvratnem prehodu pa vse vhodne parcialne odvode zmanjša za podani faktor. Shema zamrznitve oz. zmanjšanja stopnje učenja je prikazana na Sliki 3.6.

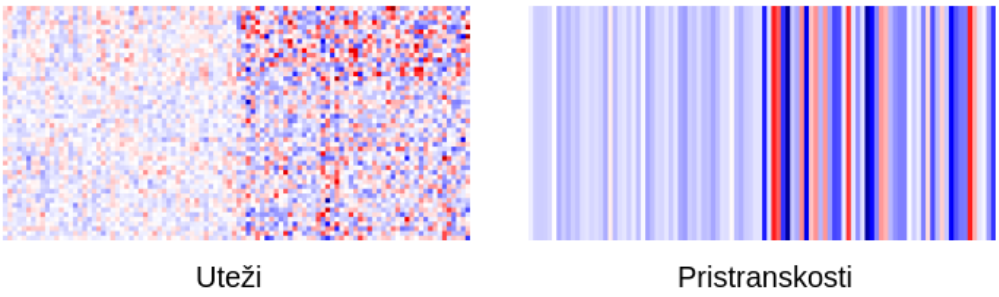
Grafi na Sliki 4.22 prikazujejo, kaj se dogaja s klasifikacijskimi točnostmi, če za vse plasti razen zadnje uporabimo zmanjšano stopnjo učenja. Oznaka



Slika 4.19: Matrika zamenjav ob koncu učenja na drugi učni množici.

na x-osi vsakega izmed grafov označuje faktor zmanjšanja. Vidimo, da je za izogib katastrofalnemu pozabljanju potrebno ogromno zmanjšanje. Klasifikacijska točnost na prvi testni množici v drugi fazi učenja pričakovano pada, ko je stopnja zmanjšanja nižja, za klasifikacijsko točnost na drugi testni množici pa velja ravno obratno. Iz grafov lahko razberemo, da zmanjšana stopnja učenja mreži omogoči, da se bolje nauči zakonitosti druge učne podmnožice, ne da bi prišlo do nemudnega katastrofalnega pozabljanja. V kolikor je stopnja zmanjšanja nižja sicer vseeno pride do pozabljanja, vendar komaj po približno 20 epohah učenja v drugi učni fazi. Po približno 10 epohah učenja je klasifikacijska točnost na celotni testni množici za približno 15% višja kot ob zamrznitvi (spodnji graf na Sliki 4.21).

Če uporabimo dovolj zmanjšano stopnjo učenja, do katastrofalnega poza-



Slika 4.20: Uteži in odmiki ob koncu učenja na drugi učni množici.

bljanja ne pride (spodnji graf na Sliki 4.22), vendar klasifikacijska točnost na celotni testni množici ne preseže tiste, ki jo dobimo z manjšim zmanjšanjem stopnje učenja in dovolj zgodnjo ustavitvijo učenja.

Preverili bomo še, kako se zmanjšana stopnja učenja obnese, če sta obe podmnožici iz iste domene.

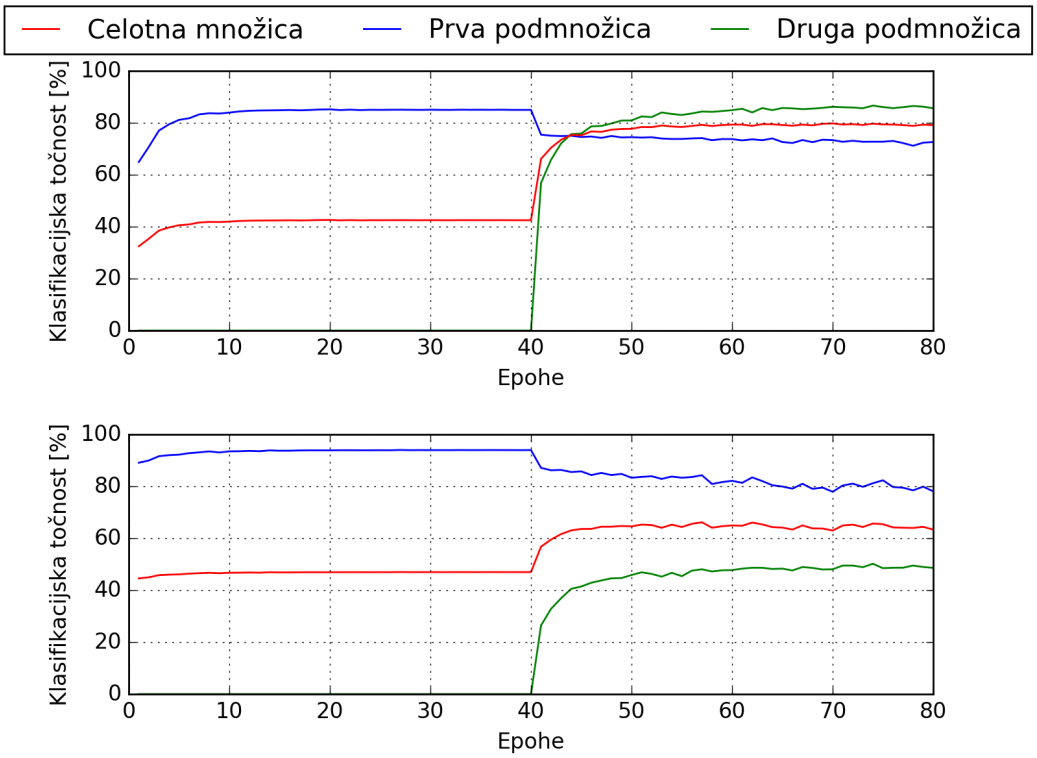
Grafi na Sliki 4.23 prikazujejo spreminjanje klasifikacijskih točnosti na obeh podmnožicah zbirke CIFAR-100 ob uporabi zmanjšanje stopnje učenja za vse plasti razen zadnje. Vidimo, da se ob ustreznem zmanjšanju stopnje učenja mreža lahko dovolj prilagodi drugi učni množici, brez da bi to porušilo klasifikacijsko točnost na prvi. Skupna klasifikacijska točnost preseže 50%, vendar moramo, v kolikor stopnje učenja ne zmanjšamo dovolj, učenje na drugi množici ustaviti dovolj zgodaj, drugače ponovno pride do katastrofalnega pozabljanja.

4.2.2 Orakelj

Izvedli smo tudi nekaj eksperimentov, kjer predpostavljamo uporabo oraklja ob testiranju nevronske mreže, kot je opisana v Razdelku 3.1.2.

4.2.2.1 Učenje z orakljem

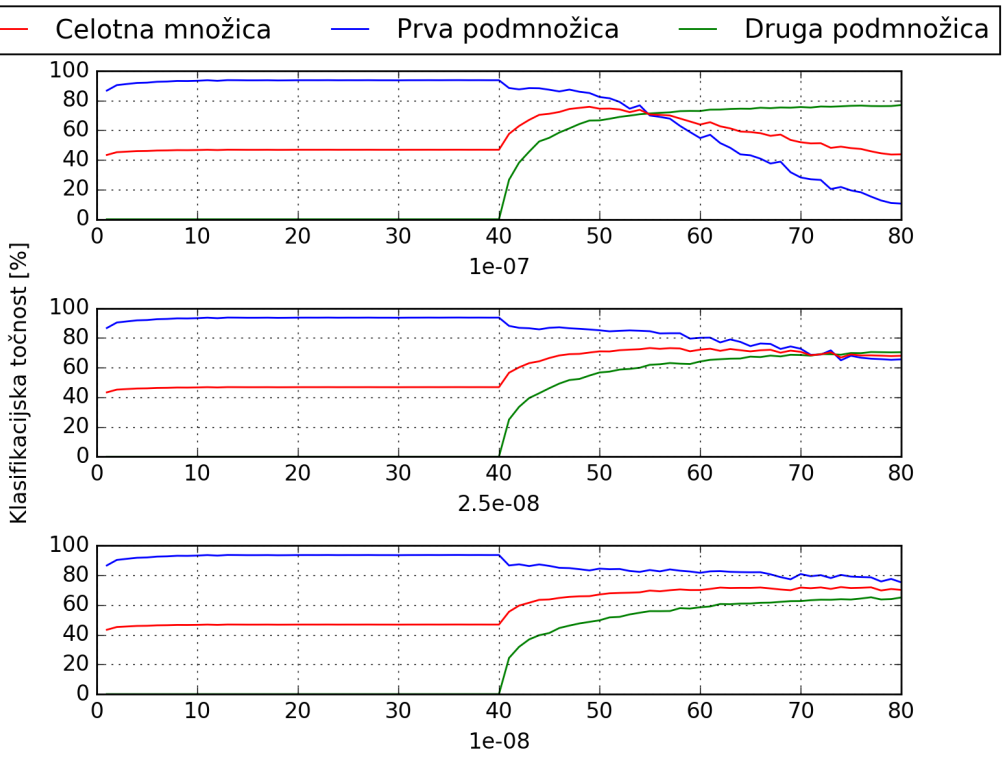
Nevronsko mrežo učimo na dveh podmnožicah podatkovne zbirke CIFAR-100, vsaka vsebuje 50 razredov. Mreža ima 50 izhodnih nevronov. Mrežo



Slika 4.21: Klasifikacijske točnosti ob učenju na Fashion-MNIST in CIFAR-10. Zgornji graf prikazuje dogajanje, če najprej učimo na CIFAR-10 in nato na Fashion-MNIST, spodnji pa obratno.

učimo za 40 epoh, z začetno stopnjo učenja 0,001, ki se vsakih 10 epoh zmanjša za desetkrat, velikost paketa je 64. Optimizacijska metoda je Adam, kriterijska funkcija je križna entropija, dosežemo približno 63% klasifikacijsko točnost na testni množici. Ko je mreža naučena shranimo, zadnjo plast in vse njene parametre, saj jih bomo potrebovali pozneje. Nato zamenjamo zadnjo plast z novo, ki ima ravno tako 50 izhodnih nevronov, preostali deli mreže se ne spreminjajo.

Nato mrežo učimo na drugih 50 razredih, vsi parametri učenja so enaki prejšnjim. Ob testiranju mreže moramo vedeti, ali primer pripada prvi ali drugi učni podmnožici in ustrezno nastaviti zadnjo plast nevronske mreže.

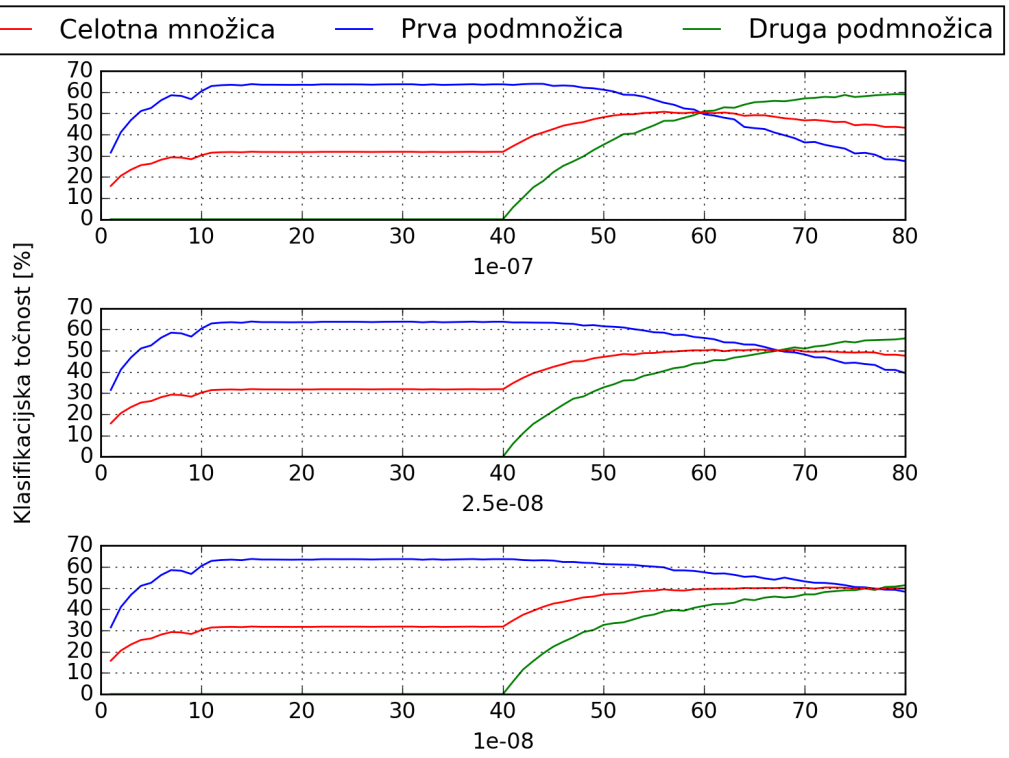


Slika 4.22: Klasifikacijske točnosti ob učenju na Fashion-MNIST in CIFAR-10 z variabilno stopnjo učenja.

Shema stanja mreže je predstavljena na Sliki 3.7.

Graf na Sliki 4.24 prikazuje spreminjanje klasifikacijske točnosti ob učenju z orakljem. Takoj opazimo, da do tako izrazitega katastrofalnega pozabljanja kot v primeru, ko oraklja ne uporabimo, ne pride. Rezultati so delno pričakovani, saj je problem ob uporabi oraklja veliko lažji. Matrika zamenjav na Sliki 4.25 predstavi, da mreža primer vedno uvrsti v enega izmed razredov iz pravilne učne podmnožice, kar je eden izmed glavnih faktorjev, ki delajo ta problem enostavnejši. Padec klasifikacijske točnosti na prvi učni množici je vseeno prisoten, saj se preostali del mreže v drugi fazi spreminja, vendar vidimo, da te spremembe nimajo tako velikega vpliva.

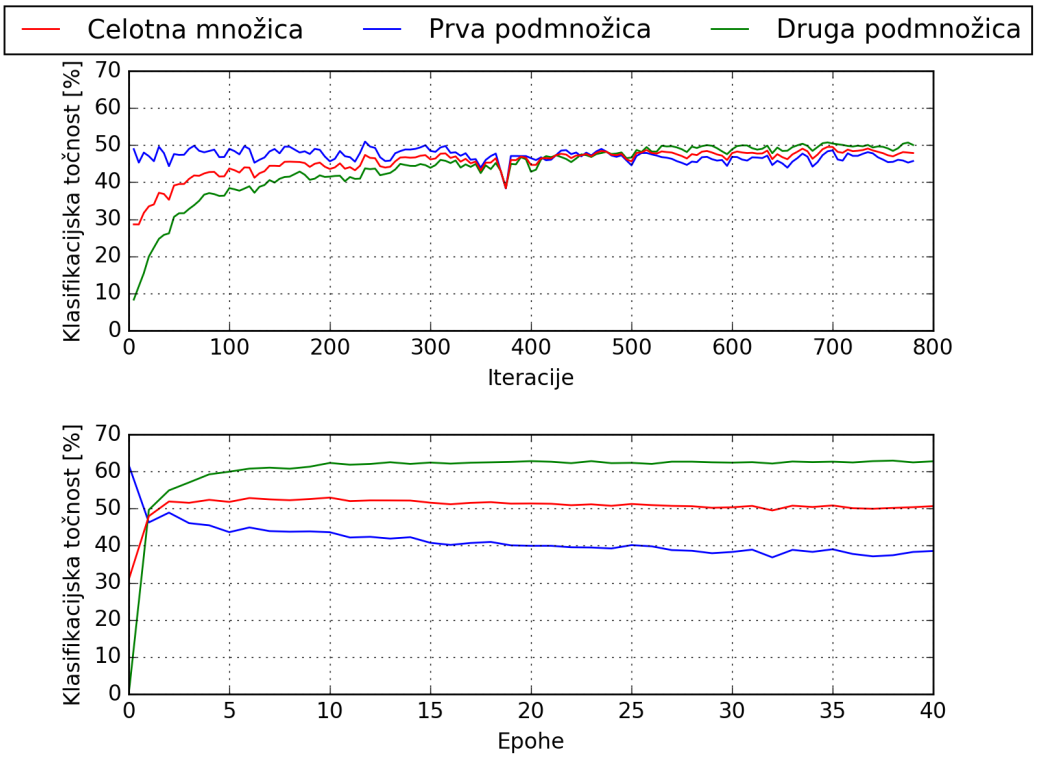
Po prvi fazi učenja mreža razmreroma dobro prepoznava primerke iz pr-



Slika 4.23: Klasifikacijske točnosti ob učenju na dveh podmnožicah CIFAR-100 z variabilno stopnjo učenja.

vih 50 razredov, ker so učni primeri iz druge podmnožice sestavljeni podobno, lahko domnevamo, da mreža v drugi fazi učenja ne potrebuje tako visoke začetne stopnje učenja. Preverili bomo, kaj se zgodi, če učenje v drugi fazi začnemo z zmanjšano stopnjo učenja. Uporabili bomo 10 krat manjšo, 0,0001.

Graf na Sliki 4.26 pokaže, da ima uporaba nižje začetne stopnje učenja zelo velik vpliv na zmanjšanje padanja klasifikacijske točnosti na prvi testni množici v drugi fazi učenja. Klasifikacijska točnost na drugi učni množici sicer raste počasneje, kot če uporabimo višjo začetno stopnjo učenja, vendar se po 40 epohah ne razlikuje bistveno od le-te. Rezultati na celotni testni množici so za približno 20% boljši, kot če uporabimo višjo začetno stopnjo



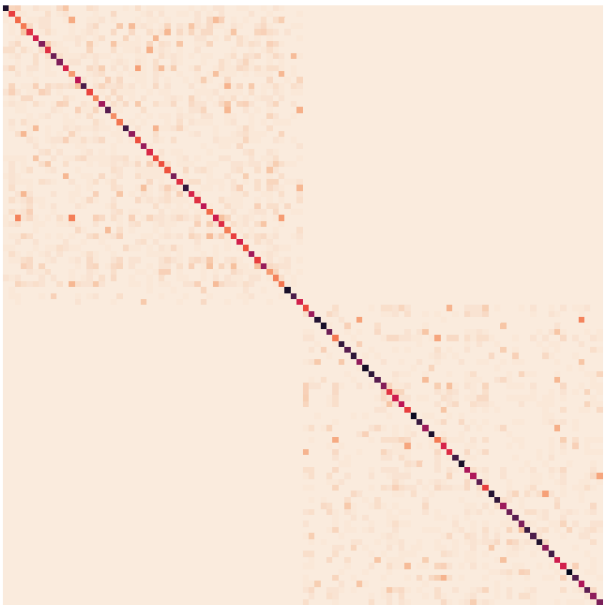
Slika 4.24: Klasifikacijske točnosti v drugi fazi učenja z orakljem. Zgornji graf prikazuje dogajanje po iteracijah znotraj prve epohe druge faze učenja, spodnji pa celotno drugo fazo.

učenja v drugi fazi (glej spodnji graf na Sliki 4.24).

Preverili bomo še, kakšen vpliv ima zamrznitev celotne mreže v drugi fazi učenja. Na grafu na Sliki 4.27 vidimo, da ob zamrznitvi mreže v drugi fazi učenja klasifikacijska točnost na drugi testni množici ne doseže dovolj visokih vrednosti, da bi bila tudi skupna klasifikacijka točnost tako visoka kot ob uporabi zmanjšane stopnje učenja.

4.2.3 Metoda MAS

Preverili bomo, kako uporaba metode MAS (Razdelek 3.1.3) vpliva na zmanjšanje katastrofalnega pozabljanja. Avtorji v članku [1] metodo uporabijo le ob

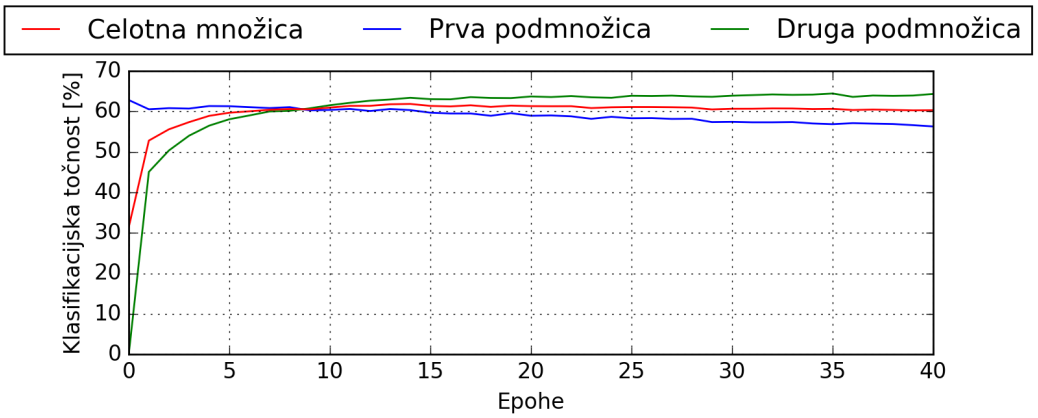


Slika 4.25: Matrika zamenjav ob koncu druge faze učenja. Vidimo, da mreža ne more prepoznati primera, kot da pripada razredu iz napačne faze učenja.

učenju z orakljem, vendar lahko postopek priredimo tako, da deluje tudi, če oraklja ne uporabljamo.

4.2.3.1 Vpliv velikosti λ

Eksperiment 4.2.2.1 bomo ponovili z uporabe metode MAS za zmanjšanje katastrofalnega pozabljanja in preverili, kako vpliva velikost hiperparametra λ na spreminjanje klasifikacijskih točnosti na obeh testnih podmnožicah. Parametri učenja se ne spreminjajo, le optimizacijsko metodo Adam moramo prilagoditi tako, da upošteva regularizacijske parametre Ω, ki jih izračunamo na prvi učni množici. Shema na Sliki 3.8 prikazuje stanje mreže v drugi fazi učenja.



Slika 4.26: Klasifikacijske točnosti v drugi fazi učenja z orakljem z zmanjšano začetno stopnjo učenja v drugi učni fazi.

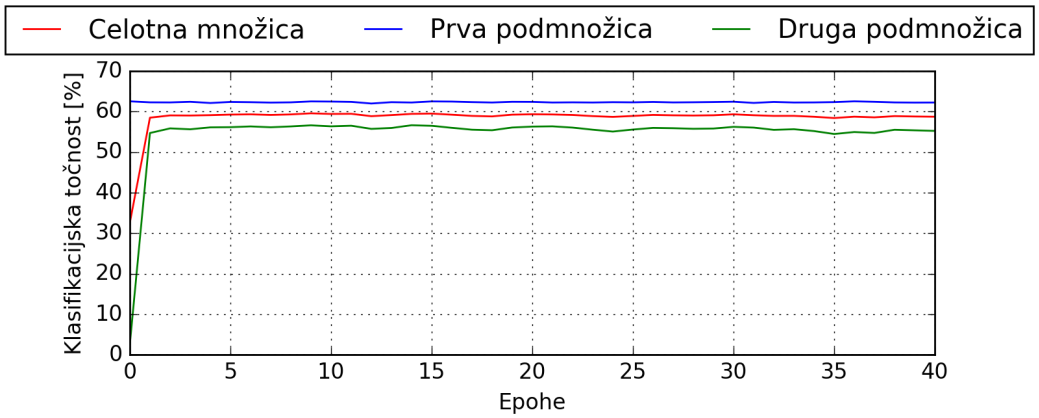
Grafi na Sliki 4.28 prikazujejo spreminjanje klasifikacije točnosti na prvi in drugi učni množici za različne vrednosti λ. Skladno z našimi pričakovanji opazimo, da klasifikacijska točnost na prvi testni množici ostaja višja z večanjem vrednosti λ, klasifikacijska točnost na drugi testni množici pa doseže višje vrednosti, ko je vrednost λ manjša.

Podobno kot v Eksperimentu 4.2.2.1 lahko tudi tukaj domnevamo, da mreža v drugi fazi učenja ne potrebuje enake stopnje učenja kot v prvi, zato bomo preverili, kaj se zgodi, če uporabimo nižjo, in sicer $0,0001$.

Grafi na Sliki 4.29 prikazujejo spreminjanje klasifikacijske točnosti za različne vrednosti λ ob uporabi nižje stopnje učenja v drugi fazi učenja. Vidimo, da se tudi tukaj katastrofalno pozabljanje močno zmanjša v primerjavi z uporabi višje začetne stopnje učenja.

4.2.3.2 MAS brez oraklja

Metodo MAS priredimo tako, da bo delovala brez oraklja. Mreža ima torej toliko izhodnih nevronov, kolikor je vseh razredov iz vseh učnih podmnožic. Pri izračunu Ω_i namesto vseh izhodnih nevronov upoštevamo le tiste, ki pripadajo do sedaj naučenim razredom. Izračun regularizacijskih parametrov



Slika 4.27: Klasifikacijske točnosti ob zamrznitvi vseh plasti razen zadnje v drugi fazi učenja z orakljem.

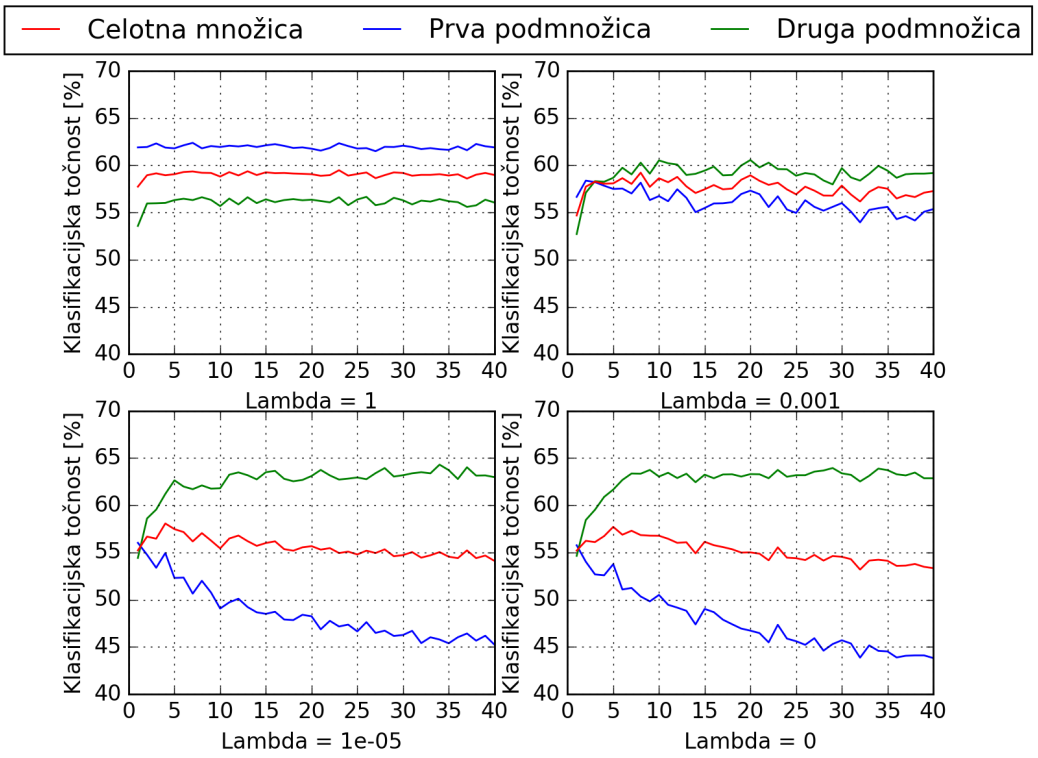
Ω ostane enak v vseh plasteh razen zadnji. V zadnji plasti bomo zamrznili parametre nevronov, ki ne pripadajo razredom iz trenutne učne podmnožice. Enak učinek bi dosegli, če bi vrednosti Ω le teh nastavili na neskončno. Vrednosti Ω parametrov, ki pripadajo nevronom v zadnji plasti in ki predstavljajo razrede iz trenutne učne podmnožice pa nastavimo na 0. Shema stanja mreže v drugi fazi učenja je prikazana na Sliki 3.9.

Grafa na Sliki 4.30 prikazujeta spreminjanje klasifikacijske točnosti ob uporabi metode MAS in odsotnosti oraklja. Na zgornjem grafu je vrednost λ enaka 1, na spodnjem pa 0,001. Vidimo, da metoda MAS ob ustrezni vrednosti λ močno zmanjša katastrofalno pozabljanje, ob vrednosti 1 klasifikacijska točnost na skupni testni množici doseže več kot 49%, kar je več, kot dobimo v Eksperimentu 4.2.1.3 (glej graf na Sliki 4.18).

4.2.4 Rezultati

Rezultate izvedenih eksperimentov smo pregledno tabelirali, ločeno glede na to, ali predpostavljamo obstoj oraklja ob uporabi mreže. Vsi rezultati so doseženi na dveh podmnožicah podatkovne zbirke CIFAR-100.

Za vse izvedene eksperimente smo zabeležili klasifikacijske točnosti na

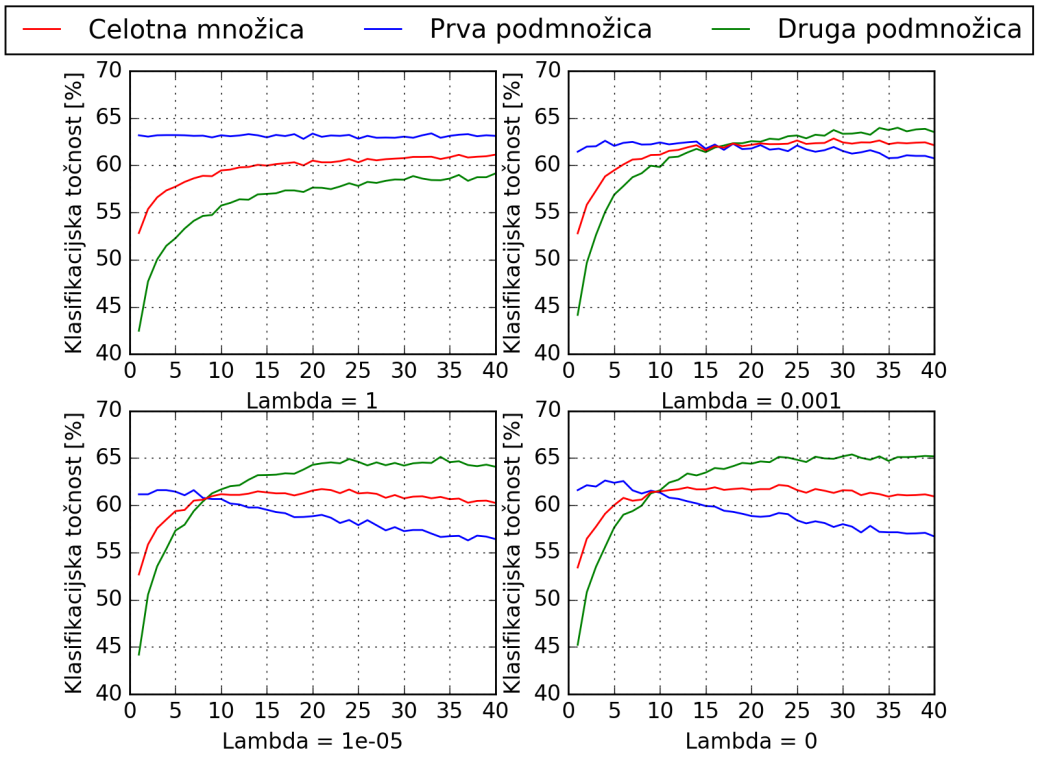


Slika 4.28: Vpliv velikosti λ na katastrofalno pozabljanje.

prvi in drugi učni podmnožici po 1 epohi učenja na drugi učni podmnožici, ob koncu druge faze učenja in ob najvišji skupni klasifikacijski točnosti. Zabeležili smo tudi končno in najvišjo skupno klasifikacijsko točnost.

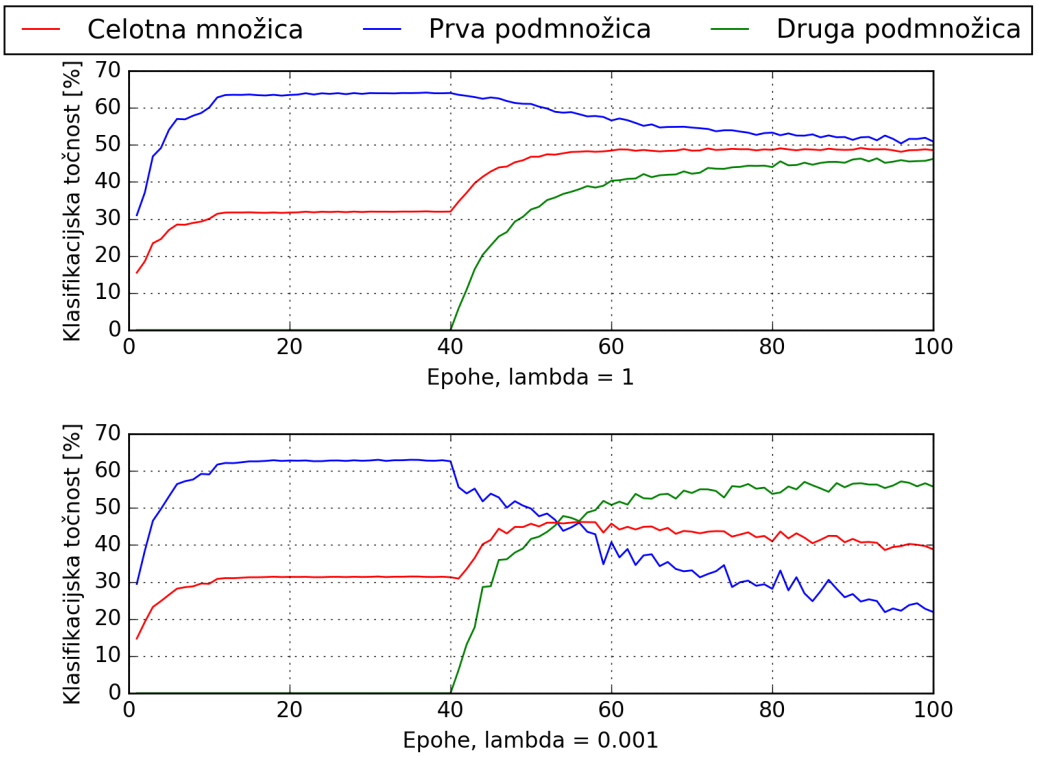
Tabela na Sliki 4.31 prikazuje klasifikacijske točnosti ob standardnem inkrementalnem učenju brez oraklja. Vidimo, da se katastrofalnemu pozabljanju praktično ne moremo izogniti, v kolikor ne zamrznemo dela nevronov v zadnji plasti, ki predstavljajo razrede iz prve učne množice. Najboljše rezultate dobimo ob uporabi variabilne stopnje učenja, rezultati metode MAS pa se jim zelo približajo.

Rezultati, ki jih dobimo ob predpostavki oraklja ob testiranju mreže, so predstavljeni na Sliki 4.32. Vidimo, da so vse klasifikacijske točnosti ob uporabi oraklja višje od najvišje brez uporabe oraklja, kar ni presenetljivo,



Slika 4.29: Vpliv velikosti λ na katastrofalno pozabljanje.

saj smo že opisali, zakaj je ta verzija problema katastrofalnega pozabljanja mnogo lažja. Metoda MAS v kombinaciji z zmanjšano stopnjo učenja v drugi fazi doseže najboljše rezultate, prav tako se katastrofalnem pozabljanju v določeni meri izognemo tudi samo z zmanjšano stopnjo učenja.



Slika 4.30: Spreminjanje klasifikacijske točnosti ob uporabi metode MAS in odsotnosti oraklja.

Eksperiment	Opombe	Klasifikacijska točnost na prvi učni množici [%]			Klasifikacijska točnost na drugi učni množici [%]			Skupna klasifikacijska točnost [%]	
		Po 1 epohi	Ob koncu	Ob najvišji skupni	Po 1 epohi	Ob koncu	Ob najvišji skupni	Končna	Najvišja
Naivni pristop	**SGD**	12,9	0,9	0,8	7,7	51,2	51,3	26,0	26,1
	SGD + zamrznitev	63,1	16,4	36,3	0,0	48,4	35,1	32,4	35,7
	Adam	1,1	0,0	0,0	53,2	61,5	63,0	30,8	31,5
	Adam + zamrznitev	4,4	0,0	2,8	53,0	53,3	55,7	26,7	29,3
Pomnenje podatkov iz prvotne množice	**Shranimo 1%**	12,2	2,8	2,8	53,8	65,9	65,9	34,4	34,4
	Shranimo 10%	38,6	24,2	35,4	48,8	61,9	54,4	43,0	44,9
	10% prvotne, 10% druge	52,9	33,7	48,5	25,0	31,3	33,8	32,5	41,2
Zamrznitev izhodne plasti	**Del izhodne plasti**	39,5	0,0	28,6	28,7	62,5	44,2	31,3	36,4
	Del izhodne in vse prejšnje plasti	60,5	43,0	49,2	12,3	48,1	45,8	45,5	47,5
Variabilna stopnja učenja	**Faktor zmanjšanja 10^8**	63,7	48,3	52,5	5,9	51,4	48,2	49,8	50,3
	Faktor zmanjšanja 10^7	63,4	27,4	55,0	5,6	58,9	46,4	43,2	50,7
MAS	**Lambda = 1**	63,5	50,9	52,0	5,9	46,2	46,3	48,6	49,2
	Lambda = 10^{-3}	55,6	21,9	46,0	6,3	55,8	46,5	38,8	46,2

Slika 4.31: Rezultati ob inkrementalnem učenju brez oraklja. Zgornjo mejo, na katero ciljamo, predstavlja 56,7%, kot lahko vidimo na zgornjem grafu na Sliki 4.3.

Eksperiment	Opombe	Klasifikacijska točnost na prvi učni množici [%]			Klasifikacijska točnost na drugi učni množici [%]			Skupna klasifikacijska točnost [%]	
		Po 1 epohi	Ob koncu	Ob najvišji skupni	Po 1 epohi	Ob koncu	Ob najvišji skupni	Končna	Najvišja
Orakelj	**Normalna SU**	46,3	38,6	43,6	49,7	62,7	62,3	50,7	53,0
	Zmanjšana SU	60,5	56,3	60,3	45,1	64,3	63,4	60,3	61,9
	Zamrznitev	62,3	62,2	62,5	54,7	55,2	56,6	58,7	59,6
MAS	**Lambda = 1**	61,9	61,9	62,4	53,6	56,0	56,3	59,0	59,4
	Lambda = 10^{-3}	56,7	55,4	58,2	52,7	59,2	60,2	57,3	59,2
	Lambda = 1 + zmanjšana SU	63,2	63,1	63,1	42,5	59,1	59,1	61,1	61,1
	Lambda = 10^{-3} + zmanjšana SU	61,4	60,7	61,9	44,1	63,5	63,7	62,1	62,8

Slika 4.32: Rezultati ob inkrementalnem učenju z orakljem. Zgornja meja, na katero ciljamo, je povprečje obeh najvišjih točnosti na spodnjih grafih na Sliki 4.3 in znaša 64,3%.

Poglavje 5

Sklep

Diplomsko delo nam omogoča podrobnejše razumevanje katastrofalnega pozabljanja pri inkrementalnem učenju globoke konvolucijske nevronske mreže. Analizirali smo, zakaj pride do katastrofalnega pozabljanja, kateri dejavniki imajo pri tem najpomembnejšo vlogo in nato na podlagi tega preizkusili nekatere ukrepe za zmanjševanje le-tega. Globlji vplogled v pojav katastrofalnega pozabljanja smo pridobili z vizualizacijo spreminjanja parametrov v drugi fazi učenja in opazovanjem matrik zamenjav.

Kot glavni problem smo prepoznali pomanjkanje oz. odsotnost učnih primerov iz prve učne podmnožice v drugi fazi učenja, ki posledično privede do konstantnega spreminjanja parametrov v smeri zmanjševanja aktivacij nevronov v zadnji plasti, ki predstavljajo razrede iz prve učne podmnožice. Videli smo, da je katastrofalno pozabljanje neizogibno, v kolikor omogočimo spreminjanje vseh parametrov mreže in v drugo učno množico ne vključimo nobenega primera iz prve. Izhajajoč iz te ugotovitve smo preverili, kakšen vpliv ima zamrznitev posameznih parametrov na katastrofalno pozabljanje.

Zamrzovanje nekaterih parametrov v zadnji plasti mreže je katastrofalno pozabljanje upočasnilo, vendar je bilo vseeno prisotno, zato smo zamrznili tudi vse preostale dele mreže in tako dobili prve spodbudne rezultate. Z zmanjšano stopnjo učenja v vseh plasteh razen zadnji in delno zamrznitvijo v zadnji nam je uspelo doseči ravnovesje med pozabljanjem na prvi podmnožici

in natačnostjo na drugi.

Preverili smo, kako na klasifikacijsko točnost vpliva prisotnost oraklja ob vrednotenju mreže in kako se le-ta spreminja s spreminjanjem stopnje učenja v drugi fazi. Implementirali smo metodo MAS za zmanjševanje katastrofalnega pozabljanja in raziskali vplive hiperparametrov metode. Metodo MAS smo nato modificirali tako, da deluje tudi brez oraklja, in jo primerjali z našimi pristopi. Ugotovili smo, da na naši nevronski mreži na primeru klasifikacije ne deluje tako dobro kot delna zamrznitev s spremenljivo stopnjo učenja.

5.1 Nadaljnje delo

Rezultati, ki jih dobimo, so sicer vzpodbudni, vendar so omejeni le na tri podatkovne zbirke, ki vsebujejo relativno majhne slike in eno konvolucijsko nevronsko mreže. V kolikor bi imeli na voljo veliko večjo računsko moč, bi izvedene eksperimente lahko ponovili na bolj obsežnih podatkovnih zbirkah, recimo ImageNet [19] ali MIT Places [22], prav tako bi lahko naše pristope preverili na kateri izmed obstoječih arhitektur, ki trenutno dosegajo najboljše klasifikacijske rezulate, npr. ResNet [8] in DenseNet [10]. Predpostavljamo lahko, da bi rezultati bili podobni, saj pri zasnovi pristopov k odpravljanju katastrofalnega pozabljanja nismo upoštevali nobenih posebnosti naše arhitekture, ampak so naši pristopi splošni in se lahko uporabijo na večini obstoječih arhitektur.

Problem katastrofalnega pozabljanja ostaja eden izmed odprtih problemov inkrementalnega oz. globokega učenja. Podrobna analiza katastrofalnega pozabljanja, predstavljena v diplomskem delu, služi kot dobra začetna točka za zasnovo novih pristopov k odpravljanju katastrofalnega pozabljanja in za razvoj inkrementalnih metod globokega učenja.

Literatura

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. *CoRR*, abs/1711.09601, 2017.

[2] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *CoRR*, abs/1710.05381, 2017.

[3] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. *CoRR*, abs/1807.09536, 2018.

[4] Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1237–1242. AAAI Press, 2011.

[5] Robert French. Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference. *Proceedings of the 16th Annual Cognitive Science Society Conference*, 08 1994.

[6] Robert M. French. Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the 13th Annual Cognitive Science Society Conference*, pages 173–178, 1991.

[7] Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. *arXiv e-prints*, page arXiv:1312.6211, Dec 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[10] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[13] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

[14] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[16] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights.

In *The European Conference on Computer Vision (ECCV)*, September 2018.

[17] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24 of *Psychology of Learning and Motivation*, pages 109 – 165. Academic Press, 1989.

[18] Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological review*, 97:285–308, 05 1990.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[20] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[21] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[22] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014.